



EXPERIENCIAS EN LA MODELACIÓN DE
LA TOMA DE DECISIONES EN LA SALUD
HUMANA, MEDIO AMBIENTE Y
DESARROLLO HUMANO

Tomo 1

Editores:

Carlos Narciso Bouza Herrera
José Félix García Rodríguez
María del Mar Rueda García
Agustin Santiago Moreno

*Experiencias en la modelación de la toma
de decisiones en la salud humana, medio
ambiente y desarrollo humano*
Tomo 1

Editores:

Carlos N. Bouza Herrera
José Félix García Rodríguez
María del Mar Rueda García
Agustín Santiago Moreno

© Universidad de la Habana (Cuba)
Universidad Juárez Autónoma de Tabasco (México)
Universidad de Granada (España)
Universidad Autónoma de Guerrero (México)

Editores:

Carlos N. Bouza Herrera
José Félix García Rodríguez
María del Mar Rueda García
Agustín Santiago Moreno

ISBN: 978-84-606-5638-8

Este libro ha sido subvencionado parcialmente por el grupo de Investigación FQM365 de la Junta de Andalucía, el proyecto MTM2012-35650 del Ministerio de Educación de España y por RIDECA.

Índice de Capítulos

CAPITULOS	AUTORES	TÍTULO	PÁGINAS
1	Cristina Calvo-Porrál , Valentín-Alejandro Martínez-Fernández y M ^a Magdalena Rodríguez-Fernández	FOOD FOR THOUGHT: THE FOOD WASTE GLOBAL ISSUE	1-23
2	Enrique Vicente-Galindo; Purificación Vicente-Galindo, M ^a Purificación Galindo-Villardón y Sergio Hernández-González	ANÁLISIS FACTORIAL, ANÁLISIS DEL IMPACTO Y TEORÍA DE RESPUESTA AL ÍTEM: APLICACIÓN AL ESTUDIO DE CALIDAD DE VIDA RELACIONADA CON LA SALUD EN PACIENTES OSTEOPORÓTICOS	24-47
3	Carlos N. Bouza Herrera y Dante Covarrubias Melgar	REVISION DE ALGUNOS PROBLEMAS ESTADISTICOS EN LA ESTIMACION DE LA BIODIVERSIDAD.	48-77
4	Yolanda Román-Montoya y Ana María Lara-Porras	FACTORES RELEVANTES EN LA INCIDENCIA DEL SIDA EN ANDALUCÍA (1992 – 2013)	78-102
5	Rilton Gonçalo Bonfim Primo and José Félix García Rodríguez	BETWEEN THE EXPLOITATION OLIGOPSONIC AND FAIR TRADEIN REVERSE LOGISTICS OF SOLID WASTE IN BRAZIL	78-102
6	Boukichou-Abdelkader, N.; Montero-Alonso, M.Á.; Muñoz-García, A. y Canário, P.N.	EXPERIMENTOS DE SIMULACIÓN: ESTIMADOR POLINOMIAL LOCAL	138-146
7	A. Santiago, J. M. Sautto, C.N. Bouza y Alfredo Montaña López	ESTIMACIÓN DEL COSTO NIVELADO POR GENERACIÓN DE ENERGÍA ELÉCTRICA EN AEROGENERADORES: EVIDENCIA EMPÍRICA DE MÉXICO	147-160
8	José Antonio Roldán-Nofuentes	ESTIMACIÓN DE PARÁMETROS DE UN TEST DIAGNÓSTICO BINARIO BAJO UN MUESTREO TRANSVERSAL	161-181
9	Luis René Marcial Castillo, Marcela Rivera Martínez, Lourdes Sandoval Solís	CLASIFICACIÓN USANDO REDES NEURONALESARTIFICIALES	182-192
10	G. Linares Fleites; E. Torres Trejo; M. de L. Sandoval Solís	SECUESTRO DE CARBONO EN SUELOS DE LA RTP 105 CUETZALAN, MÉXICO, A TRAVÉS DE TELEDETECCIÓN Y MODELOS MIXTOS	193-207

CAPITULOS	AUTORES	TÍTULO	PÁGINAS
11	L. Cabrera Hernández, L. Denoda Pérez, A. Morales Hernández, G.M. Casas Cardoso, E. González Rodríguez	MULTICLASIFICADOR PARA DIAGNOSTICAR HIPERTENSIÓN ARTERIAL INFANTIL COMBINANDO ALGORITMOS GENÉTICOS CON MEDIDAS DE DIVERSIDAD	208-225

Nombre	Afiliación
Bonfim Primo, Rilton Gonçalo	Superintendence of Solidarity Economy - SESOL of the State Secretary of Labor, Employment, Income and Sport, Brasil
Boukichou-Abdelkader, N.	Hospital Universitario La Paz. España.
Bouza Herrera, Carlos N.	Universidad de La Habana, Cuba
Cabrera-Hernández, L	Universidad Central “Marta Abreu” de Las Villas, Cuba
Calvo-Porrá, Cristina	Universidad de A Coruña (Spain)
Canário, P.N	Polytechnic Institute of Portalegre, Portugal
Casas-Cardoso, G.M.	Universidad Central “Marta Abreu” de Las Villas, Cuba
Covarrubias Melgar, Dante	Universidad Autónoma de Guerrero, México.
Denoda-Pérez, L	Universidad Central “Marta Abreu” de Las Villas, Cuba
Galindo-Villardón, Mª Purificación	Universidad de Salamanca. Salamanca, España
García Rodríguez, José Félix	Universidad Juárez Autónoma de Tabasco, Mexico
González Rodríguez. E.	Universidad Central “Marta Abreu” de Las Villas, Cuba
Hernández-González, Sergio	Universidad Veracruzana, México
Lara-Porras, Ana Maria	Universidad de Granada. España
Marcial Castillo, Luis René	Benemérita Universidad Autónoma de Puebla, México
Martínez-Fernández, Valentín-Alejandro	Universidad de A Coruña (Spain)
Montaño López, Alfredo	Universidad Autónoma de Guerrero. México.
Montero-Alonso, M.Á.	Universidad de Granada, España
Morales-Hernández. M.	Universidad Central “Marta Abreu” de Las Villas, Cuba
Muñoz-García, A.	Universidad Carlos III de Madrid, España.
Santiago Moreno, Agustín	Universidad Autónoma de Guerrero. México.
Sautto Vallejo José Maclovio	Universidad Autónoma de Guerrero. México.
Rivera Martínez, Marcela	Benemérita Universidad Autónoma de Puebla, México
Rodríguez-Fernández, Mª Magdalena	Universidad de A Coruña (Spain)
Román-Montoya, Yolanda	Universidad de Granada. España
Sandoval Solís, Lourdes	Benemérita Universidad Autónoma de Puebla, México
Vicente-Galindo ,Enrique ;	Universidad de Salamanca, España
Vicente-Galindo, Purificación	Universidad de Salamanca, España

Prólogo

Ya lo decía Huntsberger: "La palabra estadística a menudo nos trae a la mente imágenes de números apilados en grandes arreglos y tablas, de volúmenes de cifras relativas a nacimientos, muertes, impuestos, poblaciones, ingresos, deudas, créditos y así sucesivamente", pero es algo que va mucho más allá de esta simple concepción. Es para mí muy satisfactorio, como Presidente de la Asociación Mexicana de Estadística, presentar este libro a la comunidad académica sobre "Experiencias en la Modelación de la toma de decisiones en la Salud Humana, Medio Ambiente y Desarrollo Humano". El objetivo de este libro es mostrar las distintas aplicaciones que tiene la Estadística para una buena toma de decisiones, sobre todo en el ámbito de la salud humana.

Sobre Salud Humana, se presentan dos trabajos, los cuales están plasmados en el capítulo II y IV. En el primero de ellos, capítulo II, Vicente-Galindo, junto con otros colegas, comparan tres métodos estadísticos, los cuales se utilizan para simplificar los cuestionarios genéricos o específicos de Calidad de Vida Relacionada con la Salud. Presentando un ejemplo real de pacientes osteoporóticos en Atención Primaria del Centro de Salud Rural Sur de la provincia de Burgos y Centro de Salud Alfonso Sánchez Montero de Salamanca, España. En el capítulo IV, Román-Montoya y Lara-Porras, analizan datos provenientes del organismo público español Registro Andaluz de Casos de SIDA, para determinar los factores relevantes en su incidencia.

Con respecto a la temática de Medio Ambiente, este libro muestra cuatro trabajos, el primero de ellos, presentado en el primer capítulo, Calvo-Porrall y colaboradores abordan el tema del desperdicio de alimentos a nivel global y muestran que es extrema, lo cual tiene un impacto directo en dos problemas primordiales: la pobreza y el medio ambiente. El segundo de ellos, escrito en el capítulo III, por Bouza y Covarrubias, se hace una presentación de los modelos que son de uso común en el estudio de la Biodiversidad, sobre todo de los índices de biodiversidad más conocidos y utilizados. En el capítulo IX, Marcial y dos especialistas más de la Benemérita Universidad Autónoma de Puebla, describen la clasificación mediante el uso de tres modelos de redes neuronales artificiales: la multicapa, la que usa la base radial y la probabilística. El último trabajo sobre esta temática, se presenta en el capítulo X y último del libro, donde

Linares y dos autores más, nos hablan sobre el secuestro de carbono en suelos, el cual es de suma importancia para el cambio climático. Presentan y evalúan la concentración de carbono en suelo de la Zona de Cuetzan, Puebla.

Para abordar el tema de Desarrollo Humano, este libro presenta dos investigaciones. La primera de ellas está escrita en el capítulo V y la segunda en el VII. Bonfim y García, en el capítulo V hacen un análisis microeconómico del intercambio monopsónico y discuten fundamentos matemáticos clásicos del desequilibrio de precios en el sector de reciclaje de residuos sólidos en Brasil. Ya en el capítulo VII, Santiago y otros colegas, hacen una valoración monetaria de la generación de energía eléctrica usando energía eólica y presentan el caso de la generación de energía eléctrica en la zona del istmo de Tehuantepec, Oaxaca, México.

Por supuesto, no debe de faltar la parte de Modelación, aquí se presentan dos trabajos. El primero, que está en el capítulo VI, escrito por Boukichou y otros tres colaboradores españoles, realizan una simulación, con la cual pretenden estimar la tendencia de un conjunto de datos de manera no paramétrica mediante el estimador polinomial local. Ya en el capítulo VIII, Roldán-Nofuentes, hace una revisión de los parámetros de un test diagnóstico binario, el cual es una prueba médica que se aplica a un individuo para determinar si tiene o no una cierta enfermedad, además analiza las propiedades de los parámetros.

Así este libro, en mi opinión muy particular, aborda distintas aplicaciones de ésta, tan fascinante, ciencia de la Estadística.

Sergio Hernández González
Presidente de la Asociación Mexicana de Estadística

Capítulo 1.

FOOD FOR THOUGHT: THE FOOD WASTE GLOBAL ISSUE

Cristina Calvo-Porrall¹, Valentín-Alejandro Martínez-Fernández¹ y M^a Magdalena Rodríguez-Fernández¹

¹Dpto. Análisis Económico y Administración de Empresas, Universidad de A Coruña (Spain)

ABSTRACT

Between 30 and 50 percent of the total amount of food produced around the world each year -4 billion tons- never makes it on to a plate. This means that roughly one-third of food produced for human consumption is lost or wasted globally, generating huge amounts of resources used in food production used in vain, like the energy, water or land. Moreover, food is lost and wasted to a varying extent across the globe, across all stages of the food value chain and comprising all types of food. This is a tragedy in its own, given that some 870 million people suffer from chronic malnourishment. Thus, the significance of the global food waste problem is extreme, while showing a direct impact on the two most pressing issues of our time -poverty and environment-.

What causes food waste?. Extreme weather conditions and pests can lead to losses prior to harvest; machinery, transport and storage produce food waste throughout food processing. Further downstream, waste occurs due to damaged packaging, *sell by* date regulations, over purchasing and especially due to regulations and overly selective quality and aesthetic standards. However, this global issue can be tackled. On one side, for poorer countries, simply building better food storage buildings could cut down massively on food waste, along with better harvesting technology; while on the other side, industrialized countries will need to increase awareness about throwing out so much perfectly good food. In this context, the purpose of this study is to identify the causes of food waste and suggest possible ways of preventing them.

RESUMEN

Entre el 30 y el 50 por ciento del total de los alimentos producidos en el mundo cada año -4 mil millones de toneladas- nunca llega al plato. Esto significa que aproximadamente un tercio de la comida producida para consumo humano se pierde o se desperdicia a nivel global, de modo que enormes cantidades de recursos empleados en la producción de alimentos son usados en vano, como la energía, el agua o la tierra. Además, los alimentos se pierden y desperdician en distinta medida a lo largo del planeta, en todas las etapas de la cadena de valor alimentaria y para todos los tipos de alimentos. Esto supone una tragedia, dado que alrededor de 870 millones de personas sufren desnutrición crónica. Por tanto, la importancia del desperdicio de alimentos a nivel global es extrema, a la vez que muestra un impacto directo en dos de los problemas más acuciantes de nuestro tiempo -la pobreza y el medio ambiente-.

¿Cuál es la causa del desperdicio de alimentos? Las condiciones climáticas extremas y las plagas pueden conllevar pérdidas de alimentos anteriores a la cosecha; la maquinaria, el transporte y el almacenamiento también crean desperdicio de alimentos a lo largo del proceso de elaboración de alimentos. En el otro extremo de la cadena alimentaria, el desperdicio de alimentos se produce debido al empaquetado en mal estado, a la regulación de las fechas de venta de los alimentos, la compra excesiva y especialmente debido a las normas y criterios de calidad

selectivos, así como a los estándares estéticos. Sin embargo, se puede hacer frente a este problema global. Por un lado, en los países pobres, simplemente la construcción de mejores almacenes podría reducir de forma masiva el desperdicio de alimentos, así como una mejor tecnología aplicada a las cosechas; mientras que por otro lado, los países industrializados necesitarían incrementar la conciencia sobre cómo se desecha tanta comida en perfecto estado. En este contexto, el propósito de este estudio es identificar las causas del desperdicio de alimentos y sugerir posibles formas para prevenirlo.

MSC codes: 05-02

JEL codes: F6, Q2

1. THE FOOD WASTE GLOBAL ISSUE: WASTING FROM FARM TO FORK

An old traditional refrain when children did not finish the food on their plate often was “*think of all the starving children in the world*”. But, society has evolved from a culture where no food was wasted to a society where wasting food is accepted as a fact of life; highlighting that modern consumer culture includes that food produce is often wasted through retail and customer behavior (Kelleher and Robins, 2013).

In fact, the Food and Agriculture Organization of the United Nations (FAO) assumes that roughly one-third of the food produced globally for human consumption gets lost or wasted, which is about 1.3 billion tons per year; while at the same time, around 925 million people are suffering from hunger and malnutrition. Furthermore, food is wasted throughout the food supply chain, from initial agricultural production down to final household consumption, which also means a waste upon natural resources (Kranert et al., 2012).

There are numerous causes of losses and waste, being differentiated according to the different stages of the food supply chain. In developing countries, the most significant losses are concentrated at the first part of the food supply chain, primarily due to limits in the cultivation and harvesting, preserving techniques, or due to a lack of adequate transportation and storage infrastructures. However, in industrialized countries, the largest proportion of waste occurs at the final stages of the food supply chain, namely household consumption, restaurants and food services; despite there are significant losses due to aesthetic and quality standards, product regulations or production surpluses (Smil, 2010; Gustavsson et al., 2011). Consequently, the FAO estimates that food losses and food waste total about \$680 billion per year in industrialized countries and \$310 billion in developing countries (FAO 2013).

Food waste is a phenomenon that raises serious questions from a social point of view. In fact, given the problem of malnutrition that is afflicting around one billion people worldwide, the increase in food waste, appears extremely unacceptable. The lack of awareness of the scale of the food waste issue, along with its environmental, economic and social impact certainly do not assist with tackling this problem. This study highlights the food waste occurring along the entire food chain and makes an evaluation of the magnitude. Further, it identifies causes of food waste and possible ways of preventing them, along with some marketing actions which could help tackling food waste.

2. THE CONCEPTUALIZATION OF FOOD WASTE

The term *food waste* most commonly means food that was purchased but not consumed and ends up in the garbage. A first conceptualization of food waste was given by the UN Food and Agriculture Organization (FAO) including any healthy or edible substance that instead of being destined for human consumption, is wasted, lost, degraded, or consumed at every stage of the food supply chain.

There is not any standardized definition of the waste food phenomenon. So, considering of all of the phases of the food supply chain, we propose to distinguish between *food loss* and *food waste*. Both food loss and food waste only refer to products intended for human consumption, therefore excluding animal feed and non-edible parts of plants and animal products. *Food losses* refer to the decrease in edible food mass throughout the part of the supply chain that specifically leads to edible food for human consumption. Food losses take place at production, postharvest and processing stages in the food supply chain; that is, the losses that occur upstream of the food supply chain, mainly during the cultivation, harvesting, processing, preserving, and first agricultural transformation stages (Tarasuk and Eakin, 2005; Parfitt *et al.*, 2010). However food losses taking place at the end of the food chain -during industrial processing, distribution and final consumption are rather called *food waste*, which relates to retailers' and consumers' behavior. So, per definition, food waste are the masses of food lost or wasted in the part of food chains leading to edible products going to human consumption (Parfitt *et al.*, 2010).

Food waste typically occurs at the processing, storage, distribution and consumption stages of the food value chain, and is the unintended result of processes or technical limitations in storage, infrastructure, packaging or marketing; and often is the result of negligence or a conscious decision to throw food away. Consequently, any food substance, raw or cooked, which is discarded, or intended to be discarded is considered food waste, including the residues generated by the processing, handling, storage, sale, preparation, cooking and serving of foods (Griffin *et al.*, 2009; Segrè and Falasconi, 2011). On the other hand, as stated before, *food loss* refers to food that spills, spoils, incurs a reduction in quality, or otherwise gets lost before it reaches the consumer.

3. DEVELOPED AND DEVELOPING COUNTRIES: FACING DIFFERENT FOOD WASTE CAUSES.

The FAO notes that food losses in developed countries are roughly the same as in developing countries – although there is a difference in *where* all of that waste occurs. Is in the most advanced and developed countries more than 40% of food losses take place at the retail and consumer levels, while largest quantities of food is harvested from the field but never reaches the marketplace. However, more than 40% of food losses in developing countries occur at post-harvest and processing levels, (Gustavsson *et al.* 2011).

More precisely, in developing countries, wastage tends to occur primarily at the farmer-producer stage of the supply chain. The inefficient harvesting, the inadequate local transportation and the lack of infrastructure mean that produce is frequently handled inappropriately and stored under unsuitable conditions.

As the development level of a country increases, the food waste issue generally moves further up the supply chain. Despite the availability of high efficient farming practices, better transport, storage and processing facilities ensure that a larger proportion of the food produced reaches markets and consumers, the consumerism and mass marketing lead to major food wastage (Finn, 2013).

In low-income countries food is lost mostly during the early and middle stages of the food supply chain; much less food is wasted at the consumer level. As the development level of a country increases, so the food loss problem generally moves further up the supply chain with deficiencies in regional and national infrastructure having the largest impact. In developed and high-income countries food is to a significant extent wasted at the consumption stage, meaning that it is discarded even if it is still suitable for human consumption. Significant losses also occur early in the food supply chains in the industrialized regions.

As an example, it was estimated that food wastage is particularly severe in developed countries, with estimates as high as 280 to 300 kg per capita annually in Europe and in North America (Gustavsson et al., 2011); meaning that food waste and losses at the retail and consumer levels amount to an overall value of 165.6 billion dollars (Buzby and Hyman, 2012; O'Connor, 2013); but this figure does not take into account of the losses during the production and harvesting stages (Smil, 2010).

3.1. Developed countries

In richer countries, around 30 percent or more of food is discarded in processing, transportation, retailing and in consumers' kitchens. The causes of food losses and waste in medium and high-income countries mainly relate to consumer behavior, as well as to a lack of coordination between different actors in the supply chain. Farmer to retailer or distributor agreements may contribute to overproduction of crops, thereafter being wasted. In addition, food can be wasted due to quality and aesthetic standards, rejecting food products not perfect in shape or appearance. Other major reason for food waste takes place when production exceeds demand. In order to ensure delivery of agreed quantities, farmers make harvesting and production plans but end-up producing larger quantities than needed. However, in developed countries, the major contribution to food waste comes from households (Parfitt et al., 2010). At the consumer level, insufficient purchase planning and expiring best *before dates* also cause large amounts of waste, in combination with the careless attitude of consumers who can afford to waste food. As a consequence, food waste in industrialized countries can be tackled by raising awareness among food industries, retailers and consumers.

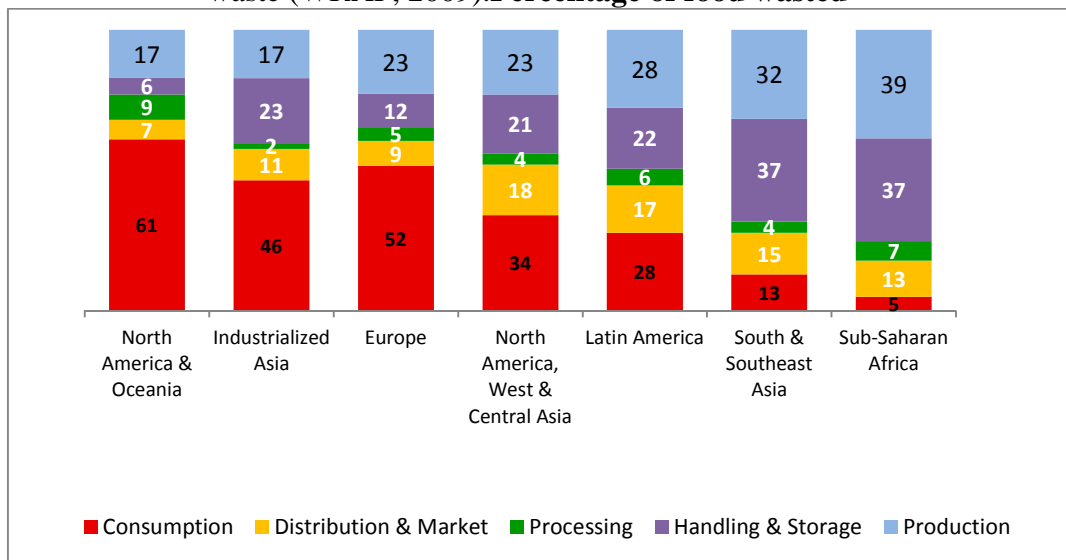
As for example, the United States Department of Agriculture (USDA) estimates that every year the United States throw away 40% of the total food intended for human consumption. So, by one recent estimate, North Americans are squandering the equivalent of \$165 billion each year by wasting food, which mostly happens in restaurants and food services (Kantor et al. 1997). Europe does a better job of curtailing waste, since recent estimates indicate that the quantity of food wasted every year amounts to 98 million tons, or 180 kg per capita.

3.2. Developing countries

The causes of food losses and waste in low-income and developing countries are mainly connected to financial, managerial and technical limitations in harvesting techniques, storage and cooling facilities, along with difficult climatic conditions, a lack of infrastructures, poor packaging and marketing systems. In poorer countries, much of the food waste occurs on the farm or soon thereafter. Depending on the crop, an estimated 20 to 35 percent of food may be lost in the farms because of poor agricultural techniques and factors such as droughts, flooding or pests; furthermore, another 10 to 15 percent is lost during processing, transport and storage given a lack of modern facilities and access to refrigeration.

Other major reason for food waste is *premature harvesting*. Poor farmers sometimes harvest crops too early due to food deficiency or the desperate need for cash. This way, the food incurs a loss in nutritional and economic value, and may get wasted if it is not suitable for consumption. However, given that many smallholder farmers in developing countries live on the margins of food, a reduction in food losses could have an immediate and significant impact on their livelihoods.

Figure 1. Differences among developing and developed countries in food waste (WRAP, 2009). **Percentage of food wasted**



As for example, a survey in India showed that at least 40 percent of all its fruit and vegetables were lost between grower and consumer and about 21 million tons of wheat annually perishes, due to lack of refrigerated transport and post-harvest cooling, along with poor infrastructures and corruption. Other example is the former Soviet Republics where cold-storage warehouses and food processing facilities were engineered and constructed in the 1930s; thus being inefficient and unsafe. As a result, Ukraine shows on average 25 to 50 percent losses in grain production.

4. THE ORIGIN OF FOOD WASTE

During the 20th century, progress in agriculture, cultivation, and the food industry allowed most developed countries to overcome the condition of food scarcity which had been prevalent before that time. Additionally, the increase in average income has given ever-larger segments of the population access to greater quantities and higher quality of food. As a result, the increasing availability and variety of food, its decreasing price, and a lower percentage of income spent on food has progressively led to greater tolerance to food waste. Furthermore, it should be remarked some global trends that had important consequences on the amount of food wasted. Firstly, the most remarkable influence is *urbanization*, which has resulted in the lengthening of the agribusiness supply chain, in order to satisfy the food requirements of the population residing in cities. The greater distance between the place of production and that of final consumption, requires the improvement of transportation, storage, and sale infrastructure in order to avoid and reduce food losses (Parfitt et al., 2010). Secondly, the change the *composition* of the human *diet*, which is associated with the increase in the available income; meaning that instead of starchy diets, consumers are increasingly demanding meat, fish and fresh products such as fruits and vegetables. This phenomenon is particularly relevant in the so-called emerging economies such as Brazil, Russia, India and China, (Lundqvist et al., 2008; Parfitt et al., 2010). Third, the increasing *globalization of commerce* and the rapid expansion and generalization of large-scale mass distribution in many developing countries; in fact, many food retailers have become the main intermediaries between food producers and consumers in Africa, Asia, and South America, thus enabling a greater diversification of diet. Finally, the demand for higher quality food products, the higher safety standards and the increase in the volume of food products marketed, clearly influence the level of food waste generated.

Table 1. Description of food waste along the value chain (Lipinski et al., 2013)

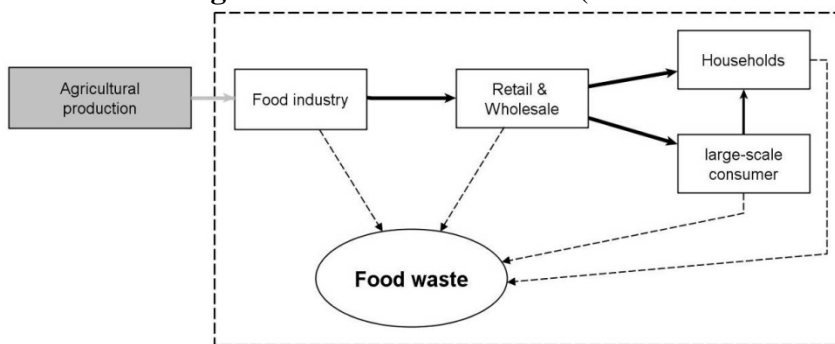
Production	Handling and storage	Processing and packaging	Distribution and marketing	Consumption
During or immediately after harvesting	After produce leaves the farm for handling, storage and transport	Industrial or domestic processing and packaging	Distribution to markets (losses at wholesale and retail markets)	Losses in the home of consumer of business (restaurants, catering...)
Stages of the value chain generating food waste				
Crops left behind in fields due to poor mechanical harvesting or drops in prices	Edible food ruined by pests	Inappropriate packaging that damages produce	Edible products spilled or damaged in market	Poor storage and stock management in houses
Fruits bruised during picking or threshing	Edible produce degraded by disease or fungus	Livestock trimming during slaughtering or industrial processing	Edible produce sorted out due to poor quality	Products sorted out due to quality
Crops damaged because of poor harvesting	Contamination and natural drying of food	Fruit or grains sorted out as not suitable for processing	Edible produce sorted out due to aesthetical standards	Food purchased but not eaten
Fish discarded during fishing operations	Fish that are spilled or degraded after landing	Fish spilled or damaged during canning or processing	Edible products expired before being purchased	Food cooked but not eaten
Crops sorted out in post-harvest for not meeting quality standards	Livestock death during transport to slaughter or not accepted for slaughter	Milk spilled during pasteurization and processing	Product discarded in supply chain	Food discarded in packaging because confusion on best before and use-by dates
			Losses caused by lack of cooling and cold appropriate storage	Food waste inconsistently treated

4.1. Production and Harvest

During the stage of the agricultural produce and farming, the main causes of waste are technical malfunctions and inefficiencies in the production processes, due to mechanical damage or spillage during harvest operation, such as threshing or fruit picking. However, the production of a sufficient amount of food is an important issue for human being, resulting in specific national and international plans to ensure the food supply, at least for developed countries. Thus, more food is produced than needed, because potentially losses due to weather incidences, diseases or pests should have consequences on food supply. This behavior is comprehensible, but implies several perspectives which may lead to food wastage (Schneider, 2008). In order to prevent food wastage in agriculture -where the weather dependency is high, forecasts have to be done to estimate the expected production; and often high amounts of edible food are wasted due to surpluses, over production, unprofitable prices as well as not marketable dimensions or weight of food products (Kranert et al., 2012).

As for example, in the United States roughly 7 percent of the produce that is grown simply gets stranded on fields each year: while some produce goes unpicked because it does not meet shape or weight standards, at times, perfectly fine crops go unpicked due to fluctuating immigration laws, which create shortages of farmworkers.

Figure 2. Areas of food waste (Source: Kranert et al., 2012)



4.2. Processing and industry

Plenty of food gets trimmed in the manufacturing and processing stage. This stage in the food chain, losses may occur when crops are sorted out if not suitable to process or during washing, peeling, slicing and boiling or during process interruptions, spillage and degradation during industrial or domestic processing. Thus, technical malfunctions in processing and refrigeration are one big factor, along with the fact that many retailers often reject produce after crops have been gathered due to quality and aesthetic standards of size, color or weight.

In first place, errors in the *production process* lead to defective or damaged food products, which can no longer be sold, while another source of waste is the spillage of products, which often occurs during the filling of liquid food (Gustavsson, et al., 2011). Additionally, production technical problems may also constitute the cause for food waste, along with damaged packaging and technical problems in labelling (Parfitt, et al., 2010; Monier, et at., 2010). Secondly, poor *storage facilities* and the lack of *infrastructure* cause postharvest food losses, mainly in developing countries. Specially, fresh produce like fruits, vegetables, meat or fish can be spoilt in hot climates, due to

lack of adequate infrastructure for transportation, storage, cooling or market (Rolle, 2006; Stuart, 2009).

Third, food products that do not comply with certain *quality* and *aesthetic characteristics* are sorted out, including for example, irregularly shaped or sized products that are not suitable for certain processes (Monier, et al., 2010). The percentage of food losses due to this reason varies depending on the type of crop and the quality standards set by regulations or requested by retailers, meaning that it is not always economically viable to collect and market certain food products. Food waste also arises during *quality assurance operations*, when food processing companies must take samples and keep them in order to check and to prove the perfect quality of their products throughout the whole shelf life; which are then disposed of (Schneider, 2009; Escaler, et al., 2011). Finally, the food loss results from *overproduction* or *poor planning* by food retailers; causes that include the *irregular demand* (Gustavsson et al., 2011; Kranert et al., 2012). Thus, it seems that it is nearly impossible to avoid food waste at this stage of the value-added chain.

4.3. Distribution and retailing

Retailing and distribution are another huge source of wasted food, since it has been estimated that food retailers loss about \$15 billion worth of unsold fruits and vegetables each year. This stage of the chain creates food waste due to spillage and degradation during handling, storage and transportation between farm and consumer, along with other causes analyzed below.

Wholesale and retail distribution waste is generally a consequence of *inappropriate ordering* and *incorrect planning of demand* for food products, resulting in large quantities of merchandise which are not sold before the expiration date or before their natural deterioration (Gustavsson, et al., 2011). Estimating demand for food products, in fact, is very complex and is influenced by multiple factors, such as climate, season, specific marketing campaigns, new product launches, promotions and even customers' attitude. Consequently, it is very difficult to predict consumers' demand; thus, leading to excess stocks (Monier, et al., 2010). According to Stenmarck (2011), the order process is one of the main problems, along with difficulties regarding the merchandise information system or in demand-based stock planning.

Other major reason leading to food waste in wholesale are the *excess stocks*. These excess stocks often result from agreements related to un-marketed food products with retailers or from orders that are cancelled at short notice (Monier et al., 2010). Food retailers focus on minimizing *stock holding*, whereas fulfilling the consumer demand, which has led to the late confirmation of order volumes as a commonplace. In fact, this confirmation of the required volumes shortly before the food products must be shipped, presents planning complexities for manufacturers and results in the creation of large amounts of food overproduction. Consequently, retailers often use different marketing strategies for food with a short shelf life in order to *reduce excess stocks* (Monier, et al., 2010). For example, in many countries, it is perfectly legal to sell food past its expiration date.

Other relevant reason for food waste at this stage is the *overproduction* and oversupply of food (Gustavsson, et al., 2011). That is, a huge variety of food products and shelves that are always completely filled, being in part responsible for wastage. Consumers

expect the shelves to be filled and a wide range of food products available in stores; thus, consequently retailers are forced to order more food products than they actually sell, causing overproduction, with more foods produced than the expected to be sold during the day in order to meet the customers' demand (Stenmarck, 2011). The large quantities on display and a wide range of food products in supply lead also to food waste, since retailers need to order a variety of food types from the same manufacturer to get beneficial prices; despite increasing the likelihood of some products reaching their sell by date before being sold, and thereby wasted. However, many stores and retailers would rather overstock their shelves and throw out the remainder than look empty; thus, food waste is seen as the cost of doing business.

Hence, *overproduction* constitutes significant cost to food companies as materials, energy and production capacity in manufacturing are wasted, given that the product no longer reaches the end customer. The underlying causes for food overproduction are mainly attributed to inefficiencies in the supply chain and to the *limited shelf life* of food products. In fact, food is mainly disposed of by retailing because it is or becomes *unsalable* or *unmarketable*, due to food products with an expired best before or use by date, along with perishable food that is often not labelled with a date, such as bread, fruits and vegetables. Therefore, a significant proportion of food waste is attributed to the *short shelf-life* of both the ingredients and products, often generated as a result of overproduction to meet retailer demands (Darlington and Rahimifar, 2006). Consequently, it may be possible to reduce the overproduction by the authorized extension of the product's *shelf life*.

Of particular significance is, however, the improper *handling* of food products, which constitutes another major reason for food losses. Quite often food products are damaged due to improper handling, and therefore can no longer be sold (Stenmarck, et al., 2011). Inadequate or improper food handling represents plenty of reasons that food is disposed of despite being undamaged: storage under insufficient temperature or lighting conditions, improper positioning of the foods products or even damaged *packaging* (Stenmarck, et al., 2011). Finally, reducing *in-store food losses* would also help in reducing food waste, which could be accomplished through better ordering and in-store inventory management.

It should be remarked that in developing countries this stage of the supply chain is characterized by a total absence or a tremendous inefficiency of wholesale distribution. The waste is attributable to the *lack of processing facilities* along with the inadequate *market systems* causing high food losses. In many situations the food processing industry does not have the capacity to process and preserve fresh produce in order to meet the demand, there are only few distributors and retail facilities providing suitable storage and sales conditions for food products, and markets are often small, overcrowded, unsanitary and lacking cooling equipment (Kader, 2005; Stuart, 2009).

4.3.1. Food waste caused by aesthetic standards

Major retailers, in meeting consumer expectations, will often reject entire crops of perfectly edible fruit and vegetables at the farm due to rigorous quality and aesthetic standards; that is, because they do not meet marketing standards concerning their physical appearance and characteristics, such as size, shape, color or weight. Although some rejected crops are used as animal feed, the high quality standards might divert

food originally aimed for human consumption to other uses (Stuart, 2009). Globally, the high appearance standards required by retailers to fresh produce, leads to generate 1.6 million tons of food waste annually in this way. One example would be the carrots quality standards required by the retailer Asda in U.K. One British farm –Poskitt Carrots- is the major supplier of carrots for this retailer. Among its production, large quantities of out-graded carrots, even having a slight bend, are sent off as animal feed. Then, all carrots pass through sensor machines in search for aesthetic defects; and thus, carrots that are not bright orange, had a blend or blemish or are broken, are swept off; since Asda requires that all carrots should be straight, so customers can peel the full length in one easy stroke (Stuart, 2009). So, finally 25 to 30% of all carrots handled by Poskitt Carrots were out-graded.

Food retailers seem convinced that consumers will not buy food which has the *wrong shape*, size, weight or appearance. However, previous research shows that consumers are willing to buy heterogeneous food produce as long as the taste is not affected, which could be done by offering consumers a broader quality range of products in the retail stores (Stuart, 2009). Consequently, consumers have the power to influence and change the *aesthetic standards*.

4.3.2. Large-scale consumers (restaurants, catering, schools, hospitals)

The main source of surplus food is overproduction due to errors in demand forecasting. Surplus food in large-scale consumers or collective catering mainly includes cooked products not yet served to customers, which is managed mainly through waste disposal, similarly to what occurs at food retailing (Garrone et al., 2014). Since food is generally regarded as the least costly resource in restaurants and catering, it is often viewed as disposable; consequently, in general terms the catering industry throws away a third of its food, as restaurants deliberately order too much produce in order to avoid running out.

Regarding the large-scale consumers, the reasons for food waste are quite similar to those of food retailing, since many reasons arise from internal organization and services for customers. In first place, in the kitchen, an *inadequate storage* and *different degrees of processing* of food products are the main causes of waste, as well as a *lack of planning* regarding the number of menus and meals to be prepared is another reason for food wastage. Since, only few kitchen chefs have recourse to data on pre-confirmed dishes, they have to rely on their experience. Secondly, much food waste arises when restaurants and caterings *do not reuse* prepared, but not yet served food. Third, and with respect to customers, waste factors include *menu selection*, *serving size* and *type* of the portions offered, given that the excessive size of food portions served will be left on the plate. However, the more customers are addressed, the more likely will be they finish their meal. This is further complicated in the case of buffet service, since buffets usually involve preparation of a larger amount of food than is necessary. Further reasons for food waste are *hygiene* and *food safety regulations* concerning the passing on and reutilization of food, as well as *lack of information*-for example, vegetarian customers want to know whether or not the meal contains meat-.In most cases, the awareness about the amount of waste generated on a daily basis in large-scale kitchens results in the decrease of food waste(Leanpath, 2006).

4.4. Household consumption

Previous studies note that overall between 30 and 50 percent of food that has been bought in developed countries is thrown away by consumers (Quester et al., 2013). Furthermore, of the food waste discarded into the garbage, approximately the 64 percent could be classified as avoidable. It seems that the culture of *abundance* and *consumer attitudes* lead to high food waste in industrialized countries, and may be one of the most important reasons for food waste in rich countries is that consumers simply can afford to waste food, while spending proportionally less income on food and less time on food preparation than previous generations (Quested et al., 2011).

Considering the food system, it should be remarked that most food decisions are made by consumers as family units, causing food waste during consumption at the household level. Surplus food at home includes both products that are bought but not consumed before the end of their shelf life, and products that are cooked but not consumed (Garrone et al., 2014). Thus, household food waste has two main causes. First is that too much food is cooked, prepared and served; thus producing the so-called *leftovers*; and second, the food that is not consumed in time, that is, food and beverages are *thrown away* because they exceeded the expiration date stated on the package, are deteriorated or no longer seem to be edible. However, household food waste reasons cannot be generalized as they depend on the product, as well as on the way of the style life of the households.

According to Glanz (2008), there is a broad range of reasons for household food waste. Consumers most often referred to the purchase of special ingredients –food products used rarely or only for certain meals-, the inadequate storage, a lack of knowledge of food stocks, food that was already spoiled before it was bought, freshness and taste - preference of freshly bought products-, products for any eventuality and trial purchases made to test new food products. Additionally, home waste arises due to the difficulties consumers face with correctly interpreting food labelling, the mistakes in the preparation or the preparation of over generous serving portions, incorrect shopping and planning and foods not being preserved appropriately.

According to Quester et al. (2013), there are two main ways of reducing the amount of food wasted in homes: by influencing consumers' behavior and attitudes, or by making changes to the food products sold and the way they are packaged. More specifically, consumers can reduce food waste at home by for example keeping fresh food produce in conditions that maximize their shelf life, planning meals in advance, checking levels of food in storage or fridge prior to shopping, making a shopping list, using the freezer to extend the shelf-life of food, portioning food like rice and pasta, using up leftovers and learning and understanding date labels on food (Quester et al., 2013).

Food wasted at consumer level is minimal in *developing countries*, since poverty and limited household income make it unacceptable to waste food. Thus, once the food gets to market, little goes to waste because food is simply considered too valuable (Finn, 2013). A contributing factor is that consumers in developing countries generally buy smaller amounts of food products at the time, often just enough for meals on the day of purchase (Stuart, 2009).

5. THE ENVIRONMENTAL, ECONOMIC AND SOCIAL IMPACT

Since roughly one third of food produced for human consumption is wasted globally - which amounts to 1.3 billion tons per year-, this fact inevitably means that huge amounts of the resources used in food production are used in vain, and that greenhouse gas emissions, caused by production of wasted food are also emissions in vain. The largest environmental benefit comes from preventing food from being wasted in the first place – consequently reducing the energy, water and resources used to grow, harvest, transport, process and sell the food-, as well as emissions associated with storage and cooking (Quested et al., 2013). Therefore, wasting food means losing not only life-supporting nourishment but also precious resources, including land, water and energy that are used in the production, processing and distribution of wasted food. So, reducing food waste can have a substantial positive environmental effect.

5.1. The environmental impact

The production of food that will not be consumed needlessly uses up natural resources and generates greenhouse gas emissions, due to the growing, processing, packaging and transporting the food to the consumer, which finally gets wasted. In fact, it should be stressed that food and drink production requires substantial inputs of water, energy and pesticides. As an example, the U.S. food waste represents the energy equivalent of 350 million of barrels of oil, enough to power the whole country for a week. More specifically, three indicators can be taken considered, namely *carbon* and *water footprint*.

On one hand, the *carbon footprint* represents greenhouse gas emissions generated during food production, comprised primarily of CO₂ generated through the use of fossil fuels, from methane derived from livestock fermentation and the emissions of nitrous oxide caused by the use of nitrogen-based fertilizers. Consequently, food waste has a large environmental impact, since avoidable food waste was estimated to contribute the equivalent of 17 million tons of CO₂ in greenhouse gas emissions annually. On the other hand, the *water footprint* is a specific indicator of the use of fresh water and is devised to convey both the actual quantities of water resources used and the way the water is used. In the case of food production, the water used in the industrial production stage is taken into account, as well as the evapotranspiration of irrigated agriculture. So, wasted food is also wasted water; since water losses accumulate as food is wasted before and after it reaches the consumer. Calculations estimate that food waste accounts for more than a quarter of total freshwater consumption globally; and the WRAP found that the water footprint of food waste was 5,400 million cubic meters, while 550 billion m³ of water is wasted globally in growing crops that never reach the consumer.

5.2. The economic and social impact

There are two main ways in which the economic impact of food waste could be assessed, referring to the production cost and the market price of food products. In first place, according to the *classic school* of economic theory, the value of a product is proportional to the resources which were necessary to produce it. Hence, the economic impact could be evaluated as the value that was lost with waste. In second place, according to the *neoclassical school* of economic theory, the value of a product does not depend on the production cost, but on its utility, represented by the market price. So, the economic impact of food waste could be estimated using the price of the individual food

products. Finally, it should be included an evaluation based on the theory of the *economics of well-being*, which estimates food waste as the impact on the usefulness of the entire society. Consequently, the economic impact should be evaluated calculating the opportunity cost of the agricultural surface used to produce the wasted goods.

The *social impact* of waste food can be tackled using the concepts of *food security* and *access to food*. Access to food is conceptualized today as a situation where “all people, at all times, have physical, social, and economic access to sufficient and nutritious food, in order to meet their dietary needs and food preferences for a healthy active life”. However, the concept of food security refers to the “availability of food nationally, in such quantities that will satisfy the energy requirements of the reference population”, meaning that along with the energy requirements, a diet should also guarantee an adequate nutrient content. Although the data concerning food supply shows an availability of food that should be sufficient to fulfill the global population’s energy requirements, there continues to be malnutrition.

6. HOW CAN FOOD WASTE BE TACKLED?

Given the scale of the problem, there are significant opportunities and solutions to be considered in order to tackle and reduce food waste, which could developed by policy makers, by farmers, food companies and consumers. In low income countries, measures should foremost have a producer perspective, by improving harvest techniques, farmer education, storage facilities and cooling chains; whereas in industrialized countries, solutions should include producer, industrial and household levels. In order to reduce food waste, the FAO (Food and Agriculture Organization), along with the UNEP (United Nations Environment Program) began a campaign targeting the aim to change the culture of food wasted by consumers and food retailers worldwide. Moreover, the European Parliament adopted a resolution in year 2012 regarding food waste, with the target of reducing food waste in EU member states by 50 percent by 2020.

Table 2. Initiatives for reducing food waste (Own elaboration from Lipinski et al., 2013)

Production	Handling and Storage	Processing and Packaging	Distribution and Marketing	Consumption
Facilitate donation of unmarketable crops	Access to low-cost handling and storage technologies	Re-think manufacturing processes	Facilitate donation of unsold foods	Facilitate donation of unsold foods from restaurants and caterers
Improve availability of agricultural extension services	Develop ethylene and microbial management of food in storage	Improve supply chain management	Change food date labeling practices	Consumer education campaigns
Improve harvesting techniques	Improve infrastructure	Develop packaging to keep food last longer	Change in-store promotions	Consumer awareness campaigns
Improve market access	Introduce refrigeration		Provide guidance on food storage and preparation	Reduce food portion sizes

6.1. Food redistribution

There are numerous initiatives worldwide aimed at the reduction and recuperation of food products that can no longer be sold but are still edible. Some of the most popular actions are known as *food redistribution* or *donation* programs, useful methods for reducing food loss and waste which allow food and retail companies to partner with charitable groups to redistribute food that would otherwise go uneaten. Following Lipinski et al. (2013) *food redistribution* means voluntarily giving away food that otherwise would be wasted to recipients -such as food banks-, which then redistribute the food to those who need it (Lipinski et al., 2013).

Thus, food redistribution applies at the production stage with crops that otherwise would go un-harvested, at the manufacturing stage with overproduced products, and at the distribution and market stage with food left unsold at stores. At the manufacturing stage, a food company might produce a surplus amount of food to meet a purchase order, but then the order is subsequently reduced by the retailer that placed it. At the market stage, surplus food might be generated when a store purchases too much of a certain food product that approaches or goes past the *sell by* date or *display until* date, printed by the manufacturer. Food is still safe to eat after these dates, as it is only an approximate measure of when a food product has reached its peak quality. However, stores are unable to sell such products in most markets due to local regulations and consumer concerns. Additionally, fresh-cooked meals at food retailers that are unsold at the end of the day are often thrown away (Lipinski et al., 2013). The major obstacles to food redistribution are related to transportation factors; since farmers and food stores with surplus food might not be physically close to food banks or charity organizations to deliver unused food.

Many organizations provide food aid through food redistribution and food donations from industry or restaurants, collecting food that would be destroyed or thrown away, and distributing it to charity. One of these organizations are the *Food Banks*. The goal of their activity is to provide surplus food from agriculture, industry -especially the food industry-, large-scale retailers and franchise restaurants; to collect food at large-scale retail centers and redistribute them to organizations dealing with aid and assistance programs. So, Food Banks collect food from a variety of sources all over the food chain, save the food in a warehouse and distribute it to poor families and individuals through local social welfare organizations, including community centers, shelters, childcare centers and senior programs. The Food Banks are able to handle surplus fruits and vegetables, which by nature have a very short life cycle requiring cold storage, and thus need a particularly rapid reception and redistribution system. More specifically, from the food industry there are thousands of companies that donate tons of products of their surplus food to the Food Banks, showing a long-standing collaboration.

Other non-profit organization dealing with food redistribution is the *Fare Share* charity association which provides quality food -collected from the food industry- to organizations that assist disadvantaged people. In some cases, the organizations recover and prepare food in community kitchens and then serve it in cafeterias, such as charity dinners, or during collective events (Segré and Gaiani, 2012). One example are the *Food Cycle* charity dinners, which organizes groups of volunteers who collect surplus food produced locally and prepares meals from it in the unused spaces of food service and restaurant kitchens, along with the development of community kitchens where meals are served at affordable prices for lunch. Other programs, such as the North

American *Rock and Wrap It Up!* recovers food that is still intact at the end of public, private or during collective events –like concerts, sporting events or business meetings– and redistributes them to Food Banks and charity organizations (Segré and Gaiani, 2012).

Other interesting initiative is the well-known *Last Minute Market*, which favors a direct relationship between supply and demand, aiming at the recovery of unsold or unmarketable food for charitable organizations. However, the *Last Minute Market* does not directly handle unsold food products, nor does it have stores, since the organization promotes the direct encounter of demand and supply, ensuring safety for all stages. The logistical and organizational system developed by *Last Minute Market* allows the completely safe recovery of all types of products, including those that fall into the category of fresh, creating a network that progressively activates the donations and pick-up system, checking nutritional, sanitation and logistical aspects. Finally, other initiatives include the use of Internet websites to market food products at discounted prices close to the recommended expiration date, such as the *Quel che c'è* initiative in Italy.

6.2. Awareness campaigns

Nowadays, the decrease in food prices coupled with an apparent abundant availability of food have led to negligence towards food and an increase in wasteful behavior (Stuart, 2009). Specifically in industrialized nations with a culture of abundance, food is relatively cheap and that plentiful nature is reinforced with every trip to attractive, fully-stocked supermarkets (Finn, 2013). Thus, there is a need to change this culture of abundance in industrialized nations, creating recognition and awareness of the *value of food*, along with the value of resources used to produce it. In fact, it is widely reported in research studies that consumers are largely unaware of the levels of food waste generated, being the lack of awareness the reason for the slow progress on reducing food waste. Hand in hand with increased awareness is the need for education in order to change food waste behavior.

Consumer attitudes and behavior play a large role in determining the amount of food that is wasted in households. Food waste at the household level occurs for a number of reasons: food leftovers on one's plate that gets thrown into the garbage or leftovers that are saved for later may be eventually thrown out, confusion over date labels such as *sell by*, *display until* and *use by*, causing some consumers to throw out food that is nutritious and safe (Lipinski et al., 2013). Moreover, some consumers do not realize the amount of waste that is actually taking place in their homes; and consequently when made aware of the staggering amounts of waste, most consumers are eager to make changes in their individual behavior.

Although changing the consumption and the food waste behavior could be difficult, communication campaigns can help influence consumer behavior at the household level. Raising awareness about food waste should be done via media campaigns in order to reach as much of the population as possible. Additional consumer information and advice to avoid waste should be communicated through internet based platforms. Since it is necessary to deal with the diverse causes of food waste one instrument could be a cooking tool to calculate proper portion serving sizes and other tool may be information on the adequate food storage (Glanz, 2008). Thus, raising food waste awareness can be effectively achieved through communication media channels such as press exposure and advertising, in order to teach consumers how to purchase, preserve, prepare, and dispose food on a more sustainable basis (Quested et al., 2013).

On one hand, food retailing can play an important role in reducing food waste at the consumption stage, because of the retailer's direct interaction with consumers. So, for example, Sainsbury's and Morrison's -the third and fourth largest grocery retailers in UK, respectively- each have created waste reduction campaigns, highlighting the food waste issue among consumers who might otherwise be uninformed, while providing them with advice for reducing waste. These two food retailers reach customers through in store displays, brochures and websites that contain recipes and information on storage and shelf lives of food products.

On the other hand, consumer education campaigns outside of retailing can also be effective in reducing food waste. One example is the ***Love Food Hate Waste*** campaign, started by the Waste and Resources Action Program (WRAP). This campaign deals with food manufacturers and retailers focused in store waste reduction initiatives, as well as with local authorities and local groups, developing activities such as interactive events, cooking demonstrations and recipe sharing gatherings, that help reduce food waste stemming from the improvement of home economics skills to unused leftovers. This initiative also comprises leaflets and newspaper advertisements providing information about how to reduce food waste and raising awareness of the environmental impact of food waste issue

Image 1. Campaign Love Food, Hate Waste directed at consumers.



However, raising awareness and provision of information alone are unlikely to change behavior across the whole population, so other methods of engagement have also been used (Quested et al., 2013). Thus, to prevent food waste, governments, municipal agencies and organizations such as the United Nations must create initiatives to help change the way people view their food, to appreciate the waste that takes place, and discourage wasteful practices. In this way, some interesting initiatives were the food waste awareness campaigns *Hate Waste* and the *This is Rubbish* developed in Great Britain or the campaign *Stop Wasting Food* in Denmark.

6.3. Re-thinking labeling

Dates provided on the packaging of food, such as *use-by*, *sell-by* and *best before*, are intended to provide consumers with information regarding the freshness and safety of foods. However, these seemingly simple dates can actually confuse consumers about how long it is safe to store food and when they should dispose uneaten food products.

Part of the confusion surrounding product labeling is that there are a number of different terms related to dates that might appear on food packages, such as, *use-by*, *sell-by* and *best before*. One date relates to the expiration of the period during which the merchandise can be displayed and sold –*sell by*-; other indicates the period during which it should be consumed –*use by*-, and still another indicates the period during which the product would be optimal –*best before*-. However, consumers often view all of these dates as being a measure of food safety –that is, a measure of whether the food could potentially cause illness when eaten-, these dates all refer to food quality or food flavor, instead of food safety. So, food that has passed its *sell by*, might be less desirable than newly-purchased food, while it is still entirely safe to eat. This misperception may lead consumers to throw away edible food they believe is no longer safe to eat. This suggests that while some of food waste may be due to safety concerns, there may be room to reduce unnecessary household food waste by clarifying the meaning of these dates and changing the way in which they are used and interpreted by consumers.

The Waste Resource Action Programme (WRAP), states that on-pack date labels and storage guidance creates consumer confusion and is a major contributing factor of food waste; thus, WRAP guidance is that customers prefer single date codes, and that *best before* is best understood. Additionally, governments could reduce food date confusion by implementing policies or providing guidance on what dates manufacturers and retailers should print on their packaging. In this way, some countries such as the UK abolish the *sell by* date which appeared on packages for all types of food products, with the goal of avoiding the waste of food products that have passed their sale dates, but were still perfectly edible.

Moreover, dates are important only to the manufacturer and the retailer could be hidden; while dates relevant to consumers could be renamed for greater clarity. More specifically, manufacturers and retailers could move to a closed date system, which would replace a *sell-by* date with a code that can be scanned by the manufacturer and retailer, but not by the consumer. Any of these actions would require a small shift in packaging, since packages would need to be slightly redesigned and would prevent consumers from misinterpreting a date on a package and throwing the food product away prematurely. Finally, retailers can reduce date-related food waste by implementing consumer education efforts, such as including in-store displays, leaflets, online guidance, or printed messages on grocery bags that define the food date labels and explain the differences between them.

6.4. Reduced portion sizes

Restaurants use larger portion sizes as a selling proposition to suggest consumers that they are receiving a bargain for the food they purchase; and consequently food portion sizes have mostly been increasing since the 70s. However, for restaurants, caterings and other food service providers, portion sizes can determine the amount of food waste that occurs within their business, since larger portions increase the likelihood that a consumer will not eat all of the food purchased (Quested et al., 2013). Thus, reducing food portion sizes in both direct –retailing- and indirect ways –restaurants and catering- can both tackle food waste and save money for food providers.

One straightforward approach for reducing this food waste would be for the restaurant to reduce portion sizes or to offer smaller portion sizes at a lower price. However, this might lead customers to feel like they are receiving less *value for money* if applied uniformly across an entire menu. This action would allow customers with smaller

appetites to order a smaller meal and presumably leave less of it behind, while also lowering preparation costs for the restaurant. Other initiative in order to reduce food waste is to remove *all-you-can-eat* restaurants and caterings, and replace them with *pay-by-weight* restaurants where the weight of the food plate determines the cost of the meal. This action might reduce food waste by giving consumers a clear economic incentive not to take more food than necessary (Lipinski et al., 2013).

7. CAN MARKETING HELP IN TACKLING FOOD WASTE?

7. 1. Food products' labelling

The way in which food is packaged, delivered, and marketed has a huge impact on disposal. Changes to packaging, labelling and merchandising can reinforce and complement the reduction of food waste (Quested et al., 2013). Research has shown that food products' *date labelling* is a determinant key factor in food disposal decisions, and that labeling of many food products can actually encourage waste. More specifically, one major cause of food waste is the strict *sell by* dates on food products, since many consumers have a poor understanding of these dates, being particularly difficult to grasp the difference between the *best by* -which relates to food quality- and *use by* -which refers to food safety-. Thus, the interpretation of the wording and the date on the food label clearly influences purchasing decisions and food disposal (Segré and Gaiani, 2012). Moreover, consumers have a tendency to select and purchase food products with a longer remaining shelf life, contributing to an increase in unsold food; and thus, to food waste (Segré and Gaiani, 2012). Consequently, some developed countries have been exploring and rethinking their use of food labels, which often encourage food retailers to toss out food long before it actually goes bad.

Similarly, many food retailers began helping consumers to use the food they buy by extending shelf life of products, providing clear and consistent date labels, improving storage and freezing guidance and also providing recipe ideas for using up leftover foods. As for example, the retailer *Morrison's* launched the campaign *Great Taste Less Waste* to help its customers waste less food, including *best kept* labelling to inform consumers how best to store food products, and promotion of effective meal planning through recipe suggestions. Other examples are the food retailers *Warburtons* and *Hovis*, which both improved on-pack labelling to make it easier for customers to know how to store their bread and when to eat it by.

7. 2. Packaging

Packaging can have a positive role in preventing food waste by extending the shelf life of food and protecting the product from damage, since damaged packaging is another reason that food is disposed of. As for example, the use of inadequate wrapping and use of wrong materials may affect healthy preservation of food and reduce the consumption period. However, a high proportion of consumers are unaware of the key role of food packaging to the preservation of food; therefore, retailers have a responsibility to communicate the benefits of packaging and inform about how food should be stored to avoid unnecessary food waste. Moreover, retailing industry might be expected to assess the magnitude of food waste related to specific packaging and processing systems and the cost-benefit relationship of alternative food products.

However, packaging will at some point go to waste, either in the supply chain or at the point of consumption, so excessive packaging is to be avoided. (Mena et al., 2011). In this way, one initiative developed by food retailers in order to meet customers' needs is the strategy of selling food products *without any packaging*. Many food retailers realized that part of their product range could be offered and sold unpackaged; and this strategy would encourage customers to buy food in accordance with their actual needs, purchasing single products instead of large packs (Kranert et al., 2012). Other interesting initiative is to develop *better packaging* in order to reduce these damages. In fact, new *packaging technologies* have provided significant advances in reducing spoilage and food contamination, thereby extending shelf-life and reducing waste. Furthermore, there exists significant potential in utilizing new materials and processes to further extend food products' integrity, such as the *polypropylene trays* for the preservation of fruits and vegetables sold as self-service, which reduces packaging and maintains the characteristics of the food product.

Some large retailers developed actions in their packaging in order to reduce food waste, considering that packaging can be crucial for the durability of food. The British retailer *Morrisons* developed the *Great Taste Less Waste* campaign, which consisted in the introduction of *best kept* stickers on fresh food to show customers the best way of preserving fresh products at home and also include practical advice in their stores. They even introduced a new packaging system for fresh and cooked products, allowing customers who need smaller food quantities to purchase a single product at a time. Moreover, the retailer *Marks & Spencer* redesigned and improved their food packaging to use less plastic and to maintain the product fresh for as long as possible; while *Warburtons* introduced a new bread loaf format designed for single consumers. Finally, the food manufacturer *Heinz* launched an innovative *fridge pack*, which could be kept in the fridge for up to five days after opening, giving consumers longer to eat the product. Hence, from a waste point of view, the decisions about what kind of packaging and how much packaging to use are critical.

7.3. Product sizes

Food manufactures produce oversized ready to eat meals and oversized food product packages (Stuart, 2009). Consequently, household food waste often reflects *excessive pack sizes* of food products especially for smaller households; thus, some particular households are at risk of producing larger quantities of waste, such as households which are single person. The reason is that while smaller pack sizes are available at a premium price, food products' price is the prevailing driver in purchasing decisions. Thus, consideration should be given to ensure that *smaller sizes* of food products are offered consistently cheaper than larger-volume food products (Defra, 2008). So, one interesting initiative to reduce food waste would consist on *buying the right amount*, which could be achieved by manufacturers and retailers offering the *right size* of pack and using promotional mechanics in order to reduce the risk of food waste. For example, the use of price reductions –25% off – as opposed to volume offers –such as 4 units for 3 euros, or buy 2 for the price of 3-. Portion size is, therefore, a significant challenge for the industry; and some manufacturers have introduced different product sizes in order to tackle this problem, like *Warburtons* and *Kingsmill* which introduced different bread loaf sizes.

Finally, restaurants should also encourage consumers to take surplus food home for later consumption -a practice which is culturally acceptable in the U.S.- The reason is that

consumers who waste food away from home do not feel a sense of ownership or responsibility about the food they leave, and the amount of food they are given is considered to be out of their control.

7.4. Food promotion offers

Food retailers' commonly used sales promotions frequently encourage consumers to purchase excessive quantities which, in the case of perishable food products inevitably generates wastage in the home. Large retailers often use different *marketing strategies* for food with a short shelf life to reduce excess stocks, which could lead the food to be shifted from retail waste to household waste (Monier, et al., 2010). Particular mention should be given in this regard to the *Buy One Get One Free* promotions, since are one of the major causes of food waste. That is, while some retail promotions offer good value for money with little impact in food waste, other promotions -like the *Buy One Get One Free* on perishable foods- can lead to an increasing food waste, by incorrect planning of food purchase, which often leads to the purchase of excessive quantities of food to take advantage of promotional offers. Moreover, the high-pressure advertising campaigns, including *bulk discounts*, also encourages consumers to buy large quantities in excess of their actual needs, which leads to substantial food wastage in the home.

In this sense, the British government also reported that *2x1 sales promotions* were among the major reasons customers purchase more food than they really need. So, some food retailers decided changing their sales promotion activities in order to reduce food waste. For example, *Sainsbury's* and *Tesco* introduced the promotion *Buy One, Get One Later*, in order to stagger purchases over time. That is, customers can buy one food item and then pick up a second one, at a later time for free, rather than having to purchase both food products at the same time. This strategy was a new spin on traditional *Buy One, Get One Free* promotions that often result in food spoiling.

Other key aspect of food waste generation also lies within retail provision, where the most compelling retail driver and competitive advantage is to maintain *stock availability* and *provision* of food commodities throughout the year. This type of supply-chain generates food waste through demand forecasting inaccuracies, product-gradings and over-production (Defra, 2008).As a consequence, the challenge for food retail management is forecasting demand and inventory management to ensure appropriate levels of stock rotation to maximize sales whilst minimizing waste. However, food demand forecasting still proves highly problematic for retailers, due to the seasonality of goods, weather patterns or even consumer trends (Defra, 2008). Finally, some retailing management practices are often culturally embedded which focuses on maintaining food products' *replenished shelves* to increase consumer satisfaction and long-term loyalty. Thus, consumers should be educated to be more tolerant of shelves that aren't stocked to full capacity right up until closing time with excessive varieties of food products.

7.5. New products and retailing concepts

The concept of *day-before bakeries* that only offer bakers' wares of the previous day can be an important way in reducing avoidable bakery food waste; along with retailers' *happy hour* concept, when products are offered at sharply reduced price levels shortly before closing time (Kranert et al., 2012). Moreover, *highlighting bargains* for food products that are nearing their *best before* date is an adequate way to reduce avoidable food waste. In this way, some food retailers give away products that have reached their

best before date to their customers or donate them to welfare services. Finally, the inclusion of *local products* in the range of goods offered by food retailers shortens transport routes, reducing the use of resources (Kranert et al., 2012).

Regarding the development of new food products, it should be remarked the initiative carried out in Denmark, by which restaurants created a partnership to produce fish snacks rich in Omega-3 from fish that was no longer saleable. Considering that more than 50% of the fish in Denmark is considered as non-edible, this experience can be an interesting solution for a product that would otherwise be discarded.

8. FOOD FOR THOUGHT

Avoidable food waste is the largest and most economical source of additional food available in our society. Consumers need to value food to a much greater degree, and there is a need to address the shortcomings of the food system in order to minimize food losses from farm to fork. Food is wasted throughout the supply chain, from initial agricultural production down to final household consumption; consequently, the involvement of all stakeholders in the food chain is needed in order to halve avoidable food waste.

While in low-income countries food is wasted mostly during the early and middle stages of the food supply chain, being the causes mainly connected to financial, managerial and technical limitations in harvesting, storage, infrastructure, packaging and marketing systems; on the other hand, in medium and high-income countries food is to a significant extent wasted at the consumption stage; being the causes of food waste mainly related to consumer behavior. Therefore, food waste in industrialized countries can be reduced by raising awareness of the scale of the global food waste issue. Moreover, some other initiatives are proposed in order to tackle food waste such as reducing pack sizes, considering the impact of in-store promotions –buy one, get one free- and clarifying date labelling.

REFERENCES

1. Buzby, J.C. and Hyman, J. (2012), “Total and per capita value of food loss in the United State”, *Food Policy* Vol. 37, pp. 561–570.
2. Cofresco Frischehalteprodukte Europa (2011), Das Wegwerfen von Lebensmitteln - Einstellungen und Verhaltensmuster. Ergebnisse Deutschland. *Save Food Study*.
3. Darlington, R. and Rahimifard, S. (2006), “A responsive demand management framework for the minimization of waste in convenience food manufacture”, *International Journal of Computer Integrated Manufacture*, Vol. 19 No. 8, pp. 751-761.
4. Defra (2008). *Evidence on the role of supplier-retailer trading relationships and practices in waste generation in the food chain*.
5. Escaler, M. and Teng, P. (2011), ‘Mind the Gap’: Reducing Waste and Losses in the Food Supply Chain. *NTS Insight*. 2011.
6. FAO (2013), “Key Findings.” *SAVE FOOD: Global Initiative on Food Losses and Waste Reduction*
7. Finn, S. (2013), “Valuing our Food: Minimizing Waste and Optimizing Resources: The Scope and Significance of the Global Food Waste Problem”, *University of Pennsylvania Scholarly Commons*.

8. Garrone, P., Melacini, M., and Perego, A. (2014), "Opening the black box of Food Waste reduction", *Food Policy*, Vol. 46, pp. 129-139.
9. Gustavsson, J., Cederberg, C., Sonesson, U., Van Otterdijk, R. and Meybeck, A. (2011). "Global Food Losses and Food Waste", Swedish Institute for Food and Biotechnology (SIK), Gothenburg, Sweden.
10. Gustavsson, Jenny, et al. 2011. Global Food Losses and Food Waste. Rome: Food and Agriculture Organization of the United Nations, 2011.
11. Glanz, R. (2008), *Causes of food waste generation in households - an empirical analysis*. Universität für Bodenkultur Wien, Institut für Abfallwirtschaft
12. Griffin, M., Sobal, J., Lyson, T. (2009), "An analysis of a community food waste stream", *Agricultural and Human Values*, Vol. 26, pp. 67–81.
13. Kader, A.A. (2005), *increasing food availability by reducing postharvest losses of fresh produce*, Proc. 5th Int. Postharvest Symp. *Acta Hort.* 682.
14. Kantor, L.S., Lipton, K., Manchester, A. and Oliveira, V. (1997), "Estimating and Addressing America's Food Losses", *Food Review*, Vol. 20 No. 1, pp. 2-12.
15. Kranert, M. (2012). Determining the quantities of food thrown away and proposals for reducing the amount of food discarded in Germany, Stuttgart University, Institute for Sanitary Engineering, Water Quality and Solid Waste Management (Germany).
16. Kelleher, M. & Robins, J. (2013). "What is waste food?", *BioCycle*, (august), pp. 36-39.
17. Leanpath (2006), Food Waste Tracking Systems. [Online] 2006. Quoted 25 January 2012.
18. Lipinski, B., Hanson, C., Lomax, J., Kitinoja, L., Waite, R., and Searchinger, T. (2013), *Reducing Food Loss and Waste*, Working Paper, World Resources Institute, Washington D.C.
19. Lundqvist, J., C. de Fraiture and D. Molden (2008), *Saving Water: From Field to Fork – Curbing Losses and Wastage in the Food Chain*. In "SIWI Policy Brief.", SIWI.
20. Mena, C., Adenso-Díaz, B., and Yurt, O. (2011), "The causes of food waste in the supplier–retailer interface: Evidences from the UK and Spain", *Resources, Conservation and Recycling* Vol. 55, pp. 648–658.
21. Monier, Véronique, et al. (2010), *Preparatory Study On Food Waste Across EU 27*. European Commission (DG ENV) Directorate C-Industry.
22. O'Connor, C. (2013), Quantification of food waste in the EU. In: 4th OECD Food Chain Network Analysis Meeting. OECD.
23. Parfitt, J., Barthel, M. & Macnaughton, S. (2010), "Food waste within food supply chains: quantification and potential for change to 2050", *Phil. Trans. R. Soc.*, Vol. 365, pp. 3065-3081.
24. Quested TE, Parry AD, Eastel S, and Swannell R. (2011), "Food and drink waste from households in the UK", *Nutrition Bulletin* Vol. 36, pp. 460–7.
25. Quested, T.E., Marsh, E., Stunell, D. and Parry, A.D. (2013), "Spaghetti soup: The complex world of food waste behaviours", *Resources, Conservation and Recycling*, Vol. 79, pp. 43-51
26. Rolle (2006). *Improving postharvest management and marketing in the Asia-Pacific region: issues and challenges*. From: Postharvest management of fruit and vegetables in the Asia-Pacific region, APO.
27. Schneider, F. (2008), "Wasting food: an insistent behavior", Proceedings of the Urban Issues & Solutions Conference, Alberta (Canada).
28. Schneider, F. (2009). Lebensmittel im Abfall - eine interdisziplinäre Herausforderung. Vortrag im Rahmen der Weiterbildungsveranstaltung „preisWERTE Lebensmittel“ am 09. Juni 2009. St. Pölten
29. Segrè, A. and L. Falasconi (2011), *Il libro nero dello spreco in Italia: il cibo*, Edizioni Ambiente.
30. Segrè, A. and S. Gaiani (2012), *Transforming Food Waste into a Resource*, RSC Publishing
31. SEPA (2008), *Svinn I livsmedelskedjan – möjligheter till minskade mängder*. Swedish Environmental Protection Agency, Bromma, Sweden

32. Smil, V. (2004), "Improving efficiency and reducing waste in our food system", *Environmental Sciences*, Vol. 1, pp. 17-26.
33. Stenmarck, Asa, et al. (2011), *Initiatives on prevention of food waste in the retail and wholesale trades*. Stockholm : Swedish Environmental Research Institute (IVL), 2011
34. Stuart, T. (2009), *Waste – uncovering the global food scandal*. Penguin Books: London.
35. Tarasuk, V. and Eakin, M. (2005), "Food assistance through "surplus" food: insights from an ethnographic study of food bank work", *Agricultural and Human Values*, Vol. 22, pp. 177–186.
36. WRAP (2009). "Household Food and Drink Waste in the UK", UK Waste Reduction Action Programme, United Kingdom.
37. WRAP (2010), "Waste arising in the supply of food and drink to households in the UK". United Kingdom.
38. WRAP (2011), "Investigation into the possible impact of promotions on food waste". Report prepared by WRAP. Banbury

Capítulo 2.

ANÁLISIS FACTORIAL, ANÁLISIS DEL IMPACTO Y TEORÍA DE RESPUESTA AL ÍTEM: APLICACIÓN AL ESTUDIO DE CALIDAD DE VIDA RELACIONADA CON LA SALUD EN PACIENTES OSTEOPORÓTICOS.

Enrique Vicente-Galindo*; Purificación Vicente-Galindo*1, M^a Purificación Galindo-Villardón2 y Sergio Hernández-González**3 * Universidad de Salamanca. Salamanca, España.
**Universidad Veracruzana. Xalapa, Veracruz, México

ABSTRACT

This paper studies the performance of three statistical methods, Factor Analysis, Impact Analysis and Item Response Theory (IRT) in the simplification of generic or specific questionnaires to measure Health Related Quality of Life (HRQL).

Along with the analysis of the theoretical characteristics of each method, the analyses are applied to QUALEFFO, a specific questionnaire to evaluate HRLQ in osteoporotic patients in Primary Care. The data analysis of 741 patients in a multicentre study, reveals that there are important discrepancies among the results of the three different methods and suggests that it is convenient not to use them separately, but jointly.

Key Words: Health Related Quality of Life, Factorial Analysis, Impact Theory, TRI, QUALEFFO, Multivariate Analysis.

RESUMEN

El trabajo consiste en realizar un estudio comparativo de tres métodos estadísticos, Análisis Factorial, Análisis del Impacto y TRI, utilizados en la simplificación de cuestionarios genéricos o específicos de Calidad de Vida Relacionada con la Salud.

Además de analizar las características teóricas, se aplican al QUALEFFO, un cuestionario específico para evaluar CV en pacientes osteoporóticos en Atención Primaria. El Análisis de los datos tomados en 741 pacientes, procedentes de un estudio multicéntrico, pone de manifiesto una clara discordancia entre los resultados de los tres métodos de Análisis y sugiere la conveniencia de utilizarlos, no individualmente, sino conjuntamente.

Palabras clave: Análisis Factorial, Impacto, TRI, CVRS, QUALEFFO

1. INTRODUCCIÓN Y OBJETIVOS

1.1. Introducción

Los primeros trabajos de psicometría se desarrollaron para medir la inteligencia. Posteriormente, la teoría psicométrica se ha aplicado a la medida de otros aspectos como los rasgos de personalidad, actitudes y creencias, rendimiento académico, y en campos relacionados con la Salud y la Calidad de Vida. Es en este campo en el que nosotros centraremos la atención ya que los datos sobre los que trabajaremos son datos de Calidad de Vida Relacionada con la Salud (CVRS) en pacientes osteoporóticos evaluados en Atención Primaria. Entendida la Calidad de Vida como una percepción subjetiva del individuo, la evaluación de la Calidad de Vida es, pues, una valoración personal; por tanto aunque hablemos de medida, la calidad de vida no se mide, se valora, pero una vez aclarado utilizaremos el término medida que es más utilizado en castellano.

Salvo en la literatura teórica, no existe una definición explícita de CV. En los trabajos empíricos, los autores no definen CV sino que se limitan a discutir las distintas formas en la que el término es definido y usado por otros. Normalmente la definición implícita viene determinada por el instrumento seleccionado para su medida.

La Osteoporosis constituye un problema de Salud Pública creciente y de primera magnitud, (ABIZANDA, 2005) tanto por su incidencia, que continua creciendo a medida que crece la esperanza de vida, como por la gravedad y el coste de sus consecuencias en forma de fractura de cadera y columna, principalmente.

Dada la trascendencia de los problemas asociados a la Osteoporosis, la detección y prevención de la misma debe ser competencia principal de la Atención Primaria de Salud. El avance en las técnicas diagnósticas y la seguridad que ofrecen los nuevos tratamientos facilitan el cambio que sin duda se está produciendo en el diagnóstico y tratamiento de la Osteoporosis, los cuales hasta hace muy poco estaban limitados a los especialistas, pero que en la actualidad han pasado a ser también competencia de los médicos de Atención Primaria.

El informe del Comité de expertos de la Unión Europea especifica en su primera recomendación que *“La Comisión Europea y los Gobiernos de los 15 Estados miembros, así como los parlamentarios y políticos nacionales, deberían adoptar explícitamente como objetivo sanitario importante la prevención de la Osteoporosis y llevar a cabo campañas de sensibilización. La prevención de la Osteoporosis debería ser prioritaria en la promoción y educación sanitaria y la formación de los profesionales de la salud”*.

1.2. OBJETIVOS

El **objetivo general** de este trabajo es comparar diferentes procedimientos utilizados para simplificar cuestionarios de calidad de vida relacionada con la salud y analizar si proporcionan resultados equivalentes.

Los **objetivos específicos** son:

1.- Aplicar la teoría del Impacto al estudio del cuestionario QUALEFFO, un cuestionario específico para evaluar calidad de Vida en pacientes osteoporóticos, con el fin de identificar los ítems más relevantes.

2.- Aplicar la teoría de Respuesta al ítem al cuestionario QUALEFFO para identificar los ítems con mayor poder discriminante.

3.- Comparar los resultados encontrados con ambos procedimientos

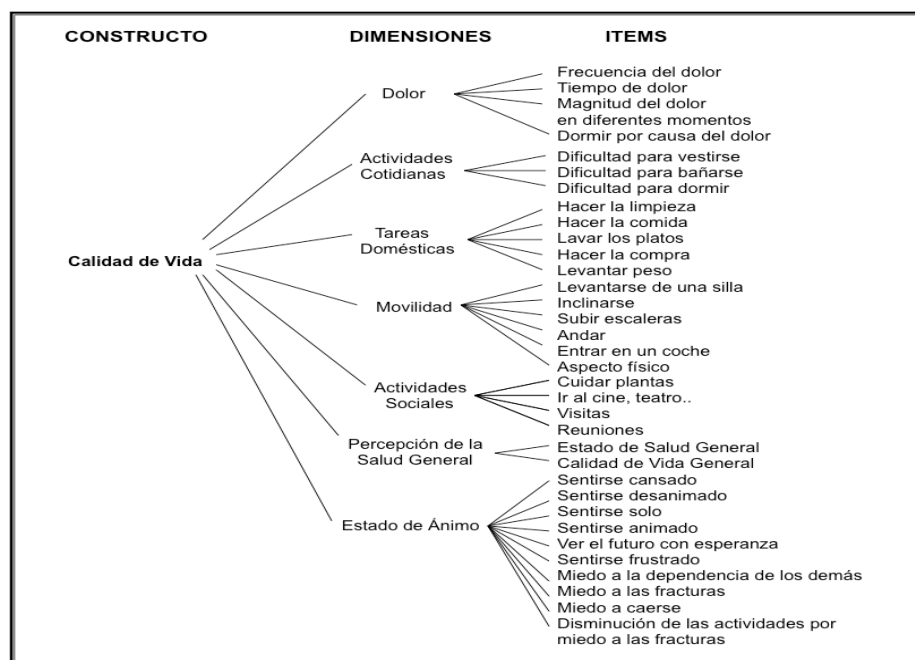
4.- Proponer una versión simplificada del cuestionario QUALEFFO basada en los resultados encontrados con ambos procedimientos.

2. MATERIAL y MÉTODOS

2.1. El cuestionario

Este trabajo se centra en el estudio psicométrico del QUALEFFO, un cuestionario específico para evaluar calidad de Vida en pacientes osteoporóticos, con el fin de estudiar si es posible eliminar algunos ítems que no aporten información relevante.

El QUALEFFO (Questionnaire of the European Foundation for Osteoporosis) (BADIA y HERDMAN, 1999), es el cuestionario específico más utilizado en medicina contemporánea para medir la Calidad de Vida en pacientes con osteoporosis, consta de 35 ítems las cuales se refieren a siete dimensiones de salud: Dolor (5 ítems), Actividades Cotidianas (3 ítems), Tareas Domésticas (5 ítems), Movilidad (6 ítems), Actividades Sociales y de Tiempo Libre (4 ítems), Percepción de la Salud General (2 ítems) y Estado de Ánimo (10 ítems). (Ver Esquema siguiente).



Estructura del Cuestionario QUALEFFO

Estas cuestiones son planteadas al paciente, y las respuestas son convertidas en puntuaciones numéricas que se combinan para proporcionar puntuaciones del dominio. En el cuestionario para medir Calidad de Vida en pacientes con Osteoporosis (QUALEFFO), las preguntas se expresan en escala politómica con cinco categorías.

El Cuestionario QUALEFFO ha sido validado en un estudio multicéntrico llevado a cabo en 7 países (LIPS y col., 1999) sobre pacientes de 55-80 años con osteoporosis clínica.

2.2. Los pacientes

Los datos con los que trabajamos están detalladamente descritos en la tesis doctoral de SÁNCHEZ BARBA, 2008; sintetizamos aquí las ideas más importantes y remitimos al lector al citado trabajo para una explicación más exhaustiva.

El diseño fue descriptivo, transversal y observacional. Se han analizado 741 pacientes de dos Centros de Atención Primaria, Centro de Salud Rural Sur de la provincia de Burgos y Centro de Salud Alfonso Sánchez Montero de Salamanca, los cuales fueron diagnosticados a partir de densitometría ósea en el hueso calcáneo utilizando métodos de ultrasonido. Los criterios de inclusión para realizar la densitometría fueron: a) tener 2 ó más factores de riesgo elevado o bien, b) tener 4 ó más factores de riesgo moderado, ó c) tener un factor de riesgo elevado y 2 ó más factores de riesgo moderado. 550 individuos de los encuestados eran mujeres, mientras que 191 eran varones. Un 18.37% de los pacientes considerados habían tenido fractura después de los 40 años; y un 9.34% tiene antecedentes familiares con fractura de caderas; el 26.20% de los 540 pacientes que contestaron a esta pregunta, tuvo menopausia prematura; el 54.94% tenía artrosis; un 7.48% tenía artritis; un 11.15% (de 260 pacientes) tiene otras enfermedades reumáticas y el 40% (de 475 pacientes) manifiesta que es el problema de los huesos el que más influye en su estado de salud. Un 32.47% de la muestra tiene menos de 60 años entre los que se encuentra que sólo hay un 12.60% de los más jóvenes (menos de 40 años); se destaca que la gran mayoría (67.53%) tienen más de 60 años. El estado civil en general apunta a que casi un 70% de la muestra está casado.

A estos pacientes se les realizó una medición ultrasónica en el calcáneo, obteniendo: el BUA que valora la densidad y estructura ósea a partir de la atenuación de la señal ultrasónica al atravesar el calcáneo, y el T-Score que nos proporciona el número de desviaciones típicas que el valor del BUA se desvía del valor medio de la población. Este último valor permitió clasificar a los pacientes estudiados en sujetos Osteoporóticos, Osteopénicos y Normales. Los puntos de corte empleados en dicha clasificación fueron de -1 DE y -2.0 DE.

Con estos puntos de corte, la muestra quedó dividida de la siguiente manera: 298 pacientes fueron considerados Osteoporóticos, 229 Osteopénicos y 192 como Normales.

2.3. Los métodos estadísticos

La evaluación científica de los instrumentos de medida de la Calidad de Vida comprende la evaluación de la fiabilidad, validez y sensibilidad al cambio (STREINER,

y NORMAN, 1995). Nosotros centraremos la atención en la validez factorial y en la capacidad discriminante.

2.3.1 Validez factorial

Aunque un instrumento debe ser fiable para que sea válido, la evidencia de fiabilidad no garantiza la evidencia para la validez; esto es, la fiabilidad es condición necesaria pero no suficiente, para la validez.

2.3.2. Análisis del impacto

El análisis del impacto, JUNIPER Y COL., 1997, consta de tres pasos. En un primer paso, se calcula la frecuencia de cada uno de los ítems del cuestionario, para así identificar los ítems más seleccionados por los sujetos. Cada uno de los 49 ítem del cuestionario OHIP, tiene 5 categorías de respuesta (0=Nunca, 1= Casi nunca, 2= a veces, 3= muchas veces, 4= siempre). Se calcula el porcentaje de individuos que aportan impacto; es decir los que puntaron en el cuestionario 2, 3, 4,5). Se re-codifica cada uno de los ítems en 0 (puntuación original 1) y 1 (resto de puntuaciones) y calcula la frecuencia de la nueva puntuación 1. En un segundo paso, calculamos la “importancia” de cada uno de los ítems del cuestionario. Consiste en calcular la puntuación media de cada uno de los ítems, sin contabilizar los que tienen puntuación NUNCA. En un tercer paso, siguiendo a ALLEN Y LOCKER, 2002 calculamos la “impacto” de cada uno de los ítems del cuestionario multiplicando la frecuencia por la importancia.

2.3.3 teoría de respuesta al ítem (tri)

En el contexto de Calidad de Vida el rasgo latente (θ) o variable latente es el “*Nivel de Calidad de Vida*” del individuo necesario para seleccionar una categoría de respuesta de cada ítem del cuestionario. De esta forma, los individuos con bajos niveles de Calidad de Vida puntuaran las categorías más altas en el ítem (generalmente).

A la probabilidad de contestar una categoría de respuesta de un ítem i a un determinado nivel de Calidad de Vida ($P_i(\theta)$), se le conoce con el nombre de función de respuesta al ítem o Curva Característica del Ítem (CCI).

La CCI muestra probabilidades cercanas a cero para valores pequeños del nivel de Calidad de Vida y probabilidades cercanas a 1 para valores elevados de la misma, produciéndose un incremento gradual de la probabilidad a medida que aumentan los valores del rasgo latente.

La probabilidad $P_i(\theta)$ depende, de dos parámetros el parámetro de discriminación y el parámetro de dificultad. (Podrían ser más pero en este trabajo se utiliza un modelo de dos parámetros).

En términos de **Calidad de Vida** el parámetro de **discriminación** a_i se interpreta como la capacidad del ítem para discriminar o diferenciar entre distintos pacientes con distinta Calidad de Vida. SÁNCHEZ BARBA, 2008. El parámetro de **dificultad** b_i se interpreta como la puntuación en la escala de Calidad de Vida necesaria que debe de tener un paciente para que sea más probable que conteste una de las categorías del ítem. Los

ítems con mayores parámetros de dificultad se corresponderán con aquellos que seleccionan los pacientes afirmativamente cuando tienen Calidad de Vida más baja

El parámetro de dificultad b_i es aquel punto de la escala latente para el cual la probabilidad de responder correctamente al ítem es, del 50%. Nos indica la posición de la curva característica en la escala latente. Los valores de la escala latente se estandarizan; los valores del parámetro de dificultad, suelen variar de -2 a +2.

El parámetro de discriminación a_i está relacionado con la pendiente de la curva característica en el punto b_i , de manera que cuanto mayor sea la pendiente de la curva, mayores serán las diferencias en las probabilidades $P_i(\theta)$ de los valores latentes próximos.

La Figura siguiente muestra una curva característica con sus parámetros fundamentales.

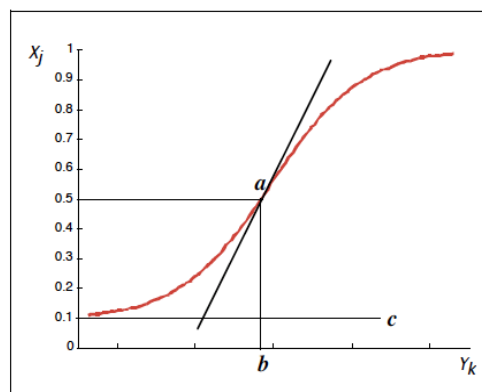


Figura 1: Representación de los parámetros de la curva característica del ítem.

El modelo logístico de dos parámetros fue propuesto por BIRNBAUM (1968).

La ecuación es la siguiente:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

donde a_i y b_i ya han sido definidos y D es una constante

Entre los distintos modelos para datos politómicos ordenados, uno de los que más atención ha recibido hasta la fecha es el modelo de Respuesta Graduada de SAMEJIMA (1969). Este modelo es una extensión del modelo logístico de dos parámetros, para el caso en que la respuesta al ítem es politómica.

COHEN y col. en 1993, definieron “los ítems de respuesta graduada” como aquellos que tienen x_i respuestas categóricas ordenadas y permiten que un individuo elija una

respuesta por ítem. Para este tipo de datos SAMEJIMA (1969) desarrollo funciones de probabilidad basadas en el modelo logístico de la TRI de dos parámetros.

El modelo se basa en las diferencias entre las funciones de respuestas categóricas, para un ítem con m respuestas categóricas, habrá $m-1$ variables binarias, por ejemplo, para un ítem de cinco categorías, la primera variable binaria está entre individuos que seleccionaron una categoría frente a las cuatro categorías superiores, la segunda está entre individuos que seleccionaron la categoría 2 o una categoría más baja frente la categoría 3 o una categoría más alta y así sucesivamente, el proceso continua hasta construir cuatro categorías. No hay necesidad de calcular la quinta variable binaria porque el ítem será puntuado como un cero cuando un individuo no seleccione ninguna categoría.

Vamos a denotar, para un ítem i , la probabilidad de responder la categoría r o superior como $P_r^{j*}(\theta)$, y para la categoría $r+1$ o superior como $P_{(r+1)}^{j*}(\theta)$.

Para el modelo de respuesta graduada logístico, la probabilidad para un ítem i de que un individuo responda a la categoría r o superior de un ítem i es:

$$P_r^{j*}(\theta) = \frac{1}{1 + e^{-Da_j(\theta - b_{(r-1)_i})}} \text{ para } r = 1, 2, 3, \dots, m, \text{ y } P_1^{j*}(\theta) = 1, P_{(m+1)}^{j*}(\theta) = 0.$$

La expresión de la *Función de respuesta categórica* para una respuesta determinada x , de un ítem i puede expresarse como:

$$P_r^j(\theta) = P_r^{j*}(\theta) - P_{(r+1)}^{j*}(\theta).$$

Sustituyendo, la expresión de la *Función de respuesta categórica* de respuesta graduada para un modelo logístico viene dado por la expresión:

$$P_r^j(\theta) = \frac{1}{1 + e^{-Da_j(\theta - b_{(r-1)_i})}} - \frac{1}{1 + e^{-Da_j(\theta - b_{r_i})}},$$

donde $P_r^j(\theta)$ es la probabilidad de que un individuo con un nivel de habilidad θ conteste a la categoría r de un ítem i ; a_j es el parámetro de discriminación del ítem i ; $b_{(r-1)_i}$ y b_{r_i} son los parámetros de dificultad para las categorías $r-1$ y r del ítem i ; y D una constante.

El número de parámetros de dificultad (b) es uno menos que el número de categorías de respuesta (si un ítem tiene cinco respuestas categóricas, los parámetros de su curva característica serán también cinco, un único parámetro de discriminación y cuatro parámetros de dificultad).

Cada parámetro de dificultad, especifica la puntuación sobre la escala latente (θ) en la que el individuo tiene un 50% de posibilidades de responder una categoría de un determinado ítem o una categoría superior.

Función de Información

La TRI utiliza el concepto de *Información del test, y de cada ítem*, para reemplazar la fiabilidad. La información es una función que varía a lo largo de la escala y tiene forma de campana, si bien la función de información del test tiende a ser mucho más variable que la de los ítems particulares.

La Función de Información es una función definida para los valores de la variable medida θ que indica, para todos los posibles valores, con qué precisión se está midiendo el test, o el ítem.

La **Función de Información del Ítem** indica que cantidad de información aporta el ítem a la medida de θ y a qué nivel aporta dicha información. La función de Información del Ítem en un nivel concreto θ , es una función en dos aspectos: a) la discriminación del ítem, de modo que a mayor pendiente, mayor información; y b) el error típico del ítem e_{θ} , de modo que a menor varianza mayor información (MARTINEZ, 1995).

Los ítems con mayor poder para discriminar tienen informaciones más altas, mientras que ítems con menor poder para discriminar tienen menor información, aunque en un rango más amplio.

3. RESULTADOS

3.1 Estructura factorial del cuestionario Qualeffo

El estudio original de validación del Cuestionario QUALEFFO (BADIA y HERDMAN, 1999) en el ámbito hospitalario define siete dimensiones de la Calidad de Vida relacionada con la Salud: *Dolor, Actividades Cotidianas, Tareas Domésticas, Movilidad, Actividades Sociales, Salud Global, Estado de Ánimo*.

El primer estudio hecho en pacientes con osteoporosis en Atención Primaria (VICENTE-GALINDO, 2003) no se encontraron exactamente esas dimensiones latentes. En este estudio se utilizó el método de componentes principales para extracción de factores y el método de rotación ortogonal tal como se hizo en el estudio original de propuesta del cuestionario con pacientes osteoporóticos.

Con los 7 primeros ejes se recoge el 72,1% de la variabilidad, explicando el primero el 43%. Según el citado estudio, el **primer factor** contiene cargas relacionadas con las dimensiones **B** (Actividades cotidianas), **C** (Tareas domésticas), **D** (Movilidad), **E** (Actividades Sociales y de Tiempo libre) y **F** (Percepción de la Salud en General), por lo que puede considerarse más relacionado con los “**aspectos y limitaciones físicas de la Calidad de Vida**”. El **segundo factor** está relacionado claramente con los ítems de la dimensión **A** (Dolor) y con el ítem B8 que también incluye la palabra dolor en su formulación ¿Tiene dificultad para dormirse a causa del dolor?. El factor 2 puede considerarse, pues, como un factor de “**Dolor**”. El **factor 3** está relacionado con algunos ítems del grupo **G** (Estado de Ánimo) concretamente con los ítems G26 (Durante los

últimos siete días, ¿se ha sentido cansado/a?), G27 (Durante los últimos siete días, ¿se ha sentido desanimado/a?) G29 (Durante los últimos siete días, ¿se ha sentido animado/a (con mucha energía?) y G30 (Durante los últimos siete días, ¿ha visto el futuro con esperanza?), y con los ítems de la dimensión **F** (Estado de Salud General); se trata, por tanto de un factor relacionado fundamentalmente con el “**estado anímico de los pacientes**”. El **factor 4** está relacionado con algunos ítems de la dimensión **G** concretamente con G32 (Durante los últimos siete días, ¿ha tenido miedo de llegar a tener que depender totalmente de los demás?), G33 (Durante los últimos siete días, ¿ha tenido miedo de tener una fractura?), G34 (Durante los últimos siete días, ¿ha tenido miedo de caerse?) y G35 (Durante los últimos siete días, ¿ha disminuido su actividad por miedo a las fracturas?), todos ellos relacionados fundamentalmente con el miedo a las fracturas. El factor podría denominarse “**miedo a las fracturas**”. Aparece una relación marginal con los ítems C13 (¿Puede levantar un objeto pesado de diez kilos de peso (por ejemplo, una caja con diez botellas de agua de un litro o un niño de 1 año) y trasladarlo por lo menos 10 metros?) y D19 (¿En qué medida le han afectado los cambios en su aspecto físico debidos a la Osteoporosis, relacionados el primero con la posibilidad de levantar peso y el segundo con la modificación del aspecto físico. El **factor 5** se relaciona solamente con los ítems G28 (Durante los últimos siete días, ¿se ha sentido sólo/a?) y G31 (Durante los últimos siete días, ¿se ha sentido frustrado/a?) correspondientes, el primero a la sensación de soledad y el segundo a la de frustración. Es entonces un factor relacionado con el “**estado de ánimo**”. El **factor 6** se relaciona con los ítems de **Actividades Cotidianas**, si bien los mismos ítems están también relacionados con el primer eje factorial. Finalmente, el **factor 7** se relaciona con A5 (Durante los últimos siete días, ¿Ha dormido mal a causa del dolor de espalda?) y B8 (¿Tiene dificultad para dormirse a causa del dolor?) que tienen como denominador común la dificultad para conciliar el sueño. Podría entenderse, pues, como un eje que representa “**trastornos del sueño**”.

De lo observado se deduce que la estructura de 7 factores supuesta a priori *no se corrobora* en estos pacientes. Solamente queda claramente definido el factor de Dolor. Todos los ítems agrupados en B, C, D y E (salvo B8) podían considerarse como un único factor que mide los aspectos de la Calidad de Vida relacionados con las limitaciones físicas, sin distinguir entre los diferentes ámbitos (el hogar o fuera del mismo) donde se producen. El factor de **Estado de Ánimo** no está claramente definido, podría desdoblarse en tres, uno propiamente relacionado con el Estado de Ánimo, compuesto por los ítems G26, G27, G29 y G30; otro relacionado con el sentimiento de soledad y frustración, que posiblemente se deba a problemas ajenos a los de la enfermedad, y otro relacionado con el miedo a las fracturas. Aparece un nuevo factor relacionado con el sueño con los ítems A5 y B8.

El factor Estado **de Salud General** no queda claramente definido ya que sus cargas se reparten entre el nuevo factor de consecuencias físicas y el de estado de ánimo, poniendo de manifiesto, probablemente, que el individuo cuando responde a la pregunta global sobre Calidad de Vida tiene en cuenta (aunque no sea conscientemente) las dos dimensiones, la física y la psíquica. Lo anterior se sintetiza en la tabla siguiente:

Factor	Dimensiones
1	Salud Física + Ocio / T. Libre + Acti Sociales
2	Dolor
3	Soledad / Frustración
4	Miedo a las Fracturas
5	Estado de Ánimo
6	Cuidado Personal
7	Trastornos del sueño

Tabla 2: Factores latentes encontrados en el conjunto de pacientes Osteoporóticos de Atención Primaria (VICENTE, 2003)

Según afirma VICENTE-GALINDO, 2003, otros autores ya habían obtenido resultados similares a los encontrados en Atención Primaria, en pacientes con Osteoporosis, si bien el estudio se hizo con la versión con 41 ítems del cuestionario QUALEFFO (MURRELL y col, 2001).

Aunque los autores que crearon el cuestionario, y aquellos que lo han validado en España hablan de 7 (6 + CV general) dimensiones de la CVRS al evaluar pacientes Osteoporóticos, en realidad eso es matizable. Veamos el razonamiento de VICENTE-GALINDO, 2003:

- Los 5 primeros ítems se refieren a **Dolor**, pero el ítem A5 pregunta sobre si *¿ha dormido bien a causa del dolor?*, luego una respuesta negativa evidencia problemas de dolor pero también trastorno del sueño.
- Los ítems que van del B6 al D19 son los que en otros cuestionarios componen una dimensión general que los autores identifican como Salud “física” o funcionamiento físico, aunque cada bloque, el B, el C y el D, tienen sus connotaciones particulares.
- En los ítems B etiquetados como “**Actividades cotidianas**” se engloba bañarse y vestirse y también dificultad para dormir a causa del dolor. Es obvio que dormir es una actividad cotidiana, pero también traduce trastornos del sueño debidos al dolor, luego no sería sorprendente encontrar asociación entre el ítem A5 y el B8, de la misma manera que podría encontrarse la relación entre el A5 y el resto de los ítems que comienzan con A y entre el ítem B8 y el resto de los B.

VICENTE-GALINDO, 2003 analizó, utilizando SPSS, las estructuras factoriales no solo en el colectivo completo sino también en Osteoporóticos en Osteopénicos y en Normales por separado, dado que la prevalencia de osteoporosis y osteopenia en pacientes que acuden al Hospital para un examen radiológico es mucho mayor de la encontrada en Atención Primaria en un Screening. No encontró las estructuras latentes hipotetizadas en el estudio original en el que se validó el cuestionario en España (las 7 que hemos señalado más arriba). Estudió además las congruencias entre las estructuras

con el Metodo de AHMAVAARA¹.La componente asociada a la Salud Física no aparece claramente definida en el grupo de Osteopénicos. Para el caso de la Osteopenia la componente de Limitaciones Físicas de Osteoporosis se correlaciona con las componentes 1, 6 y en menor medida la 4. Las componentes están, cada una de ellas relacionadas con aspectos parciales de la Salud Física. Para el caso de los pacientes normales las correlaciones son incluso más bajas, lo que indica que las Limitaciones Físicas no se manifiestan de la misma manera. La explicación de los resultados puede ser que los pacientes que no padecen todavía la enfermedad no tienen las Limitaciones Físicas de los enfermos, aunque en los Osteopénicos se adivinan ya algunas de ellas.La componente de Dolor si está presente en todos los grupos, como ya vimos en las descripciones individuales. Como podemos observar en la tabla las correlaciones entre las componentes etiquetadas como “Dolor” en los tres grupos son bastantes altas.La componente etiquetada como Soledad y Frustración del grupo “Osteoporosis” no aparece claramente en el grupo de “Osteopenia” y aparece correlacionada con la etiquetada como “Estado de Ánimo” en el grupo “Normal”.La componente que habíamos denominado “Estado de Ánimo” en el grupo de “Osteoporosis” aparece también en el grupo de “Osteopenia” (la correlación es alta, 0.9051) pero no aparece tan clara en el grupo de los “Normales”, ya que la correlación con la componente correspondiente de la misma denominación es más baja. La componente de problemas relacionados con “Trastornos del Sueño” aparece solamente en el grupo de pacientes con Osteoporosis. Lo mismo ocurre con la componente de Actividades Cotidianas, aunque hay una cierta relación con algunas de las componentes del grupo “Osteopenia”.La componente relacionada con el Miedo a las Fracturas aparece en los tres grupos.En conclusión, no podemos afirmar que haya una invarianza estructural de la Calidad de Vida, en los distintos estadios de la enfermedad. La estructura latente de la Calidad de Vida en los pacientes Osteoporóticos no es la misma que en los pacientes con Osteopenia. En los pacientes con la enfermedad más severa se produce un deterioro en las capacidades físicas que no aparece globalmente en los Osteopénicos; ahí aparecen bloques de ítems correlacionados pero no todos ellos constituyen una única dimensión latente de Deterioro Físico. La componente de Dolor aparece en todos los grupos, lo cual es lógico teniendo en cuenta que los pacientes no han sido seleccionados aleatoriamente sino que son pacientes con factores de riesgo en los cuales es muy probable que los dolores se den, por ésta o por otra razón. El miedo a las fracturas podría ser consecuencia también de los problemas asociados con el dolor.Aunque VICENTE-GALINDO no encontró las 7 dimensiones postuladas en el instrumento QUALEFFO y tampoco encontró invarianza estructural, lo que si es cierto es que hay ciertas dimensiones que si se detectan de forma invariante en los tres grupos; concretamente la relativa al Dolor y la relativa al Miedo a las Fracturas. Ciertas manifestaciones del Estado de Ánimo positivo y negativo, y ciertas manifestaciones de

¹El procedimiento de AHMAVAARA, 1954, pensado para comparar pares de estructuras, consiste simplemente en el cálculo de los cosenos de los ángulos entre las componentes de dos grupos, los cuales pueden entenderse como coeficientes de correlación entre las componentes.

Salud Física, también están presentes en los tres grupos pero con correlaciones más bajas.

3.2. RESULTADOS DEL ANÁLISIS DEL IMPACTO

Tal como ya hemos señalado, la definición de Calidad de Vida es problemática. Se han usado numerosas aproximaciones pero ninguna ha sido universalmente aceptada. Ya hemos señalado más arriba que, salvo en la literatura teórica, no existe una definición explícita de CV. En los trabajos empíricos, los autores no definen CV sino que se limitan a discutir las distintas formas en la que el término es definido y usado por otros. **Normalmente la definición implícita viene determinada por el instrumento seleccionado para su medida.** De ahí la trascendencia de encontrar instrumentos psicométricamente robustos, fáciles de administrar, fáciles de comprender, y que no necesiten mucho tiempo para ser contestados.

Ya hemos podido comprobar como las estructuras latentes que teóricamente conformar un instrumento (cuestionario) no siempre se encuentran en el colectivo que estudiamos², a veces los cuestionarios son demasiados largos y/o confusos, otras veces los ítems están relacionados con varias dimensiones, y un largo etc. Este hecho ha propiciado que muchos trabajos de investigación estén dirigidos a la simplificación de los cuestionarios para evitar ambigüedad y sobre todo para ganar en eficacia.

Son muchos los intentos en este sentido; uno de los más actuales y menos conocidos es el análisis del impacto de los ítems. Tal como ya señalamos en el apartado de Métodos, en el análisis del impacto, JUNIPER Y COL., (1997), se calcula en primer lugar la frecuencia de cada uno de los ítems del cuestionario, para así identificar los ítems más seleccionados por los sujetos. Cada uno de los 49 ítem del cuestionario OHIP, tiene 5 categorías de respuesta (0=Nunca, 1= Casi nunca, 2= a veces, 3= muchas veces, 4= siempre). Se calcula el porcentaje de individuos que aportan impacto; es decir los que punturan en el cuestionario 2, 3, 4, 5).

En nuestro caso, se re-codifica cada uno de los ítems en 0 (puntuación original 1) y 1 (resto de puntuaciones) y calcula la frecuencia de la nueva puntuación 1.

Si analizamos la tabla siguiente podemos observar, teniendo en cuenta que se trata de un cuestionario específico para evaluar Calidad de Vida en pacientes osteoporóticos, que los ítems que más frecuentemente han respondido afirmativamente (aunque no todos con el mismo nivel), **para la dimensión A(Dolor)** han sido los ítems A1 (Durante los últimos siete días, ¿con qué frecuencia ha tenido dolor de espalda?), A2 (Durante los últimos siete días, ¿cuánto tiempo le ha durado el dolor de espalda durante el día?) y A3 (Durante los últimos siete días, ¿cómo ha sido el dolor de espalda en el peor de los casos?), los tres representan presencia de dolor en los pacientes durante los últimos 7 días. Sin embargo el ítem A5 (Durante los últimos siete días, ¿Ha dormido mal a causa del dolor de espalda?) representa un síntoma mucho menos frecuente, solo un 28.2% dice haber dormido mal alguna noche (puede que haya sido una noche, o dos, casi todas o todas), pero alguna ha dormido mal.**Para la dimensión B(Actividades Cotidianas),**

²En realidad cuando el cuestionario tiene muchos ítems es muy probable que esto suceda

los tres ítems representan síntomas actuales pero que son mucho menos frecuentes que los de la escala anterior. Solo el ítem B8 (¿Tiene dificultad para dormirse a causa del dolor?) tiene asociado un porcentaje que supera el 34%, Este resultado es concordante con el obtenido para el dolor-sueño en la dimensión latente anterior. Según los valores obtenidos para los ítems, B6 (Tiene dificultad para vestirse) y el B7 (Tiene dificultad para bañarse o ducharse) solo un 23% dice tener esos problemas en la actualidad. Para la **dimensión C (Tareas Domésticas)** los ítems C13 (¿Puede levantar un objeto pesado de diez kilos de peso (por ejemplo, una caja con diez botellas de agua de un litro o un niño de 1 año) y trasladarlo por lo menos 10 metros?) y C9 (¿Puede hacer la limpieza?) representan los síntomas más frecuentes (actualmente). Los demás ítems de esa dimensión presentan porcentajes mucho más bajos. Ver detalles en la tabla siguiente. Para la dimensión **D (Movilidad)**, los ítems D17 (¿Puede andar 100 metros?) y D16 (¿Puede subir las escaleras entre dos pisos?) son los más frecuentes en el colectivo estudiado. Para la dimensión **E (Actividades Sociales y de Tiempo libre)**, son los ítems E20 (¿Puede cuidar de las plantas de su terraza o jardín?) y E21 (¿Puede ir al cine, al teatro, a un restaurante, etc.?) los más frecuentes con 45.3% y 20.80 % respectivamente. Los otros dos ítems de esta escala toman valores similares a los del ítem E21. Para la escala **G (Estado de ánimo)**, los ítems G29 (Durante los últimos siete días, ¿se ha sentido animada (con mucha energía)?) y G30 (Durante los últimos siete días, ¿ha visto el futuro con esperanza?) son los que presentan porcentajes más altos de 78 y 73% respectivamente.

En un segundo paso, calculamos la “importancia” de cada uno de los ítems del cuestionario. Siguiendo a ALLEN Y LOCKER, 2002, consiste en calcular la puntuación media de cada uno de los ítems, sin contabilizar los que tienen puntuación 1=nunca. En la tabla 4 se presentan los ítems con mayor importancia: A1 y A5 de la escala A, B6 y B7 de la escala B, C13 y C12 de la escala C, D15 y D16 de la escala D, de la escala E presentan todos los ítems valores de importancia prácticamente idénticos y de la G, los ítems G33 y G34. Más detalles pueden consultarse en la tabla 4

En un tercer paso calculamos la “impacto” de cada uno de los ítems del cuestionario. Siguiendo a ALLEN Y LOCKER, 2002, se calcula multiplicando la frecuencia por la importancia. El autor sugiere seleccionar como ítems para conformar el cuestionario simplificado, los ítems de mayor impacto. Para que estén representadas todas las dimensiones latentes del constructo el autor sugiere hacer esa elección, no de forma absoluta sino eligiendo los dos ítems con más impacto en cada dimensión. El análisis se ha llevado a cabo en EXCEL.

Los marcados en amarillo son los que, según esta teoría, deberían ser los seleccionados para el cuestionario simplificado. Obsérvese que no siempre los más frecuentes, ni los de mayor importancia son los que tienen el máximo impacto. ***La selección se hace según el valor del impacto.*** Estos ítems son:

Dimensión A: A1 y A2; Dimensión B: B7 y B8; Dimensión C: C12 y C13; Dimensión D: D15 y D16; Dimensión E: E20 y E21; Dimensión G: G29 y G30

Para más detalles consultar la tabla 3 en la página siguiente, en la cual tenemos los ítems y sus impactos correspondientes.

Item	Frecuencia	Importancia	Impacto
A1	52,500	3,910	205,275
A2	52,600	3,350	176,210
A3	52,500	3,170	166,425
A4	47,200	2,800	132,160
A5	28,200	3,540	99,828
B6	23,300	2,660	61,978
B7	23,100	3,070	70,917
B8	34,500	2,650	91,425
C9	40,100	2,840	113,884
C10	19,300	2,700	52,110
C11	21,100	2,740	57,814
C12	37,400	3,150	117,810
C13	54,800	3,570	195,636
D14	31,300	2,590	81,067
D15	48,700	3,090	150,483
D16	51,800	2,720	140,896
D17	52,900	2,430	128,547
D18	37,500	2,630	98,625
D19	41,600	2,770	115,232
E20	45,300	2,970	134,541
E21	20,800	3,050	63,440
E22	19,800	3,040	60,192
E23	19,800	3,000	59,400
F24	96,100	3,230	310,403
F25	95,800	2,980	285,484
G26	70,200	3,460	242,892
G27	60,700	3,370	204,559
G28	37,200	3,360	124,992
G29	83,700	3,550	297,135
G29 invertida	78,000	3,420	266,760
G30	82,900	3,710	307,559
G30 invertida	73,000	3,440	251,120
G31	44,400	3,260	144,744
G32	49,300	3,670	180,931
G33	45,700	3,710	169,547
G34	50,100	3,780	189,378
G35	30,400	3,640	110,656

Tabla 3: Frecuencia, importancia e impacto de los ítems del QUALEFFO

3.3. Resultados de la tri aplicada al análisis del cuestionario Qualeffo

El estudio de la posible simplificación del cuestionario en base a su capacidad informativa, aplicando la teoría de respuesta al ítem, ha sido realizado por SÁNCHEZ-BARBA en 2008. Estudió el comportamiento de cada una de las dimensiones que conforma el cuestionario. Para la estimación de los parámetros, el ajuste de los modelos, así como las curvas características y de información del ítem, se utilizó el programa

MULTILOG (THISSEN, 1991), ya que tiene implementado el modelo de Respuesta Graduada de SAMEJIMA (1969). Para la estimación de los parámetros utilizó el procedimiento de estimación marginal de máxima verosimilitud y se evaluó el ajuste de cada uno de los ítems con la prueba estadística de χ^2 .

ESCALA A: DOLOR

En el siguiente gráfico tomado de SANCHEZ-BARBA con permiso de la autora, vienen representadas la **función de información** (línea azul) y el **error estándar** (línea roja) para la dimensión del *Dolor*.

La curva del error estándar asociado a la **función de información** pone de manifiesto que esta dimensión *Dolor* no aporta información para los pacientes que tienen buena (o muy buena) Calidad de Vida pero si, para los que tienen mala Calidad de Vida en sus diferentes grados; es decir si hay dolor la Calidad de Vida es mala (tanto más cuanto más dolor se percibe) pero si no hay dolor no está asegurada una buena Calidad de Vida percibida ya que el paciente puede presentar otra característica que es la que motiva su percepción.

La máxima información de la dimensión del *Dolor* alcanza el valor **44.94** (Ver Gráfico 2), lo que significa que la información media por ítem estaría en torno a **8.99**. El punto de rasgo latente en el que se alcanza la información máxima está en torno al 0.60; lo que significa que para la dimensión del Dolor, el punto de máxima información se alcanza para un nivel de Calidad de Vida “regular”.

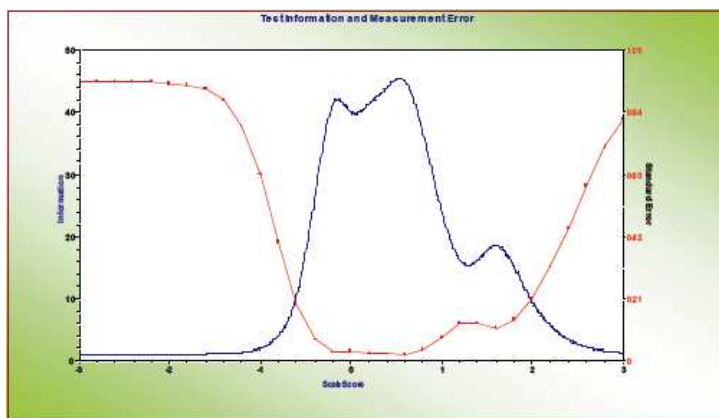


Figura 2.- Función de Información y Error Estándar de la dimensión Dolor

Tomada de SANCHEZ_BARBA, 2008

Si comparamos la información aportada por cada ítem a la dimensión del *Dolor* con la información media antes señalada, se puede observar que los ítems **A4** (*Durante los últimos siete días, ¿cómo ha sido el dolor de espalda en otros momentos?*) y **A5** (*Durante los últimos siete días, ¿ha dormido mal a causa del dolor de espalda?*) aportan menos información que el resto de los ítems (Ver figura 3). La información que

aportan los ítems **A1** (*Durante los últimos siete días, ¿con qué frecuencia ha tenido dolor de espalda?*), **A2** (*Durante los últimos siete días, ¿cuánto tiempo le ha durado el dolor de espalda durante el día?*) y **A3** (*Durante los últimos siete días, ¿cómo ha sido el dolor de espalda en el peor de los casos?*) está por encima de la información media esperable (Ver Tabla 4).

El ítem que aporta mayor información a la hora de evaluar la Calidad de Vida percibida por el paciente (en la dimensión del *Dolor*) es el ítem **A1** (*Durante los últimos siete días, ¿con qué frecuencia ha tenido dolor de espalda?*) y lo hace sobre todo en las categorías de respuesta centrales.

DOLOR		
ÍTEMS	Valor Máxima Información	Punto Máxima Información
A1. Durante los últimos siete días, ¿con qué frecuencia ha tenido dolor de espalda?	16.88	0.4
A2. Durante los últimos siete días, ¿cuánto tiempo le ha durado el dolor de espalda durante el día?	11.64	0.6
A3. Durante los últimos siete días, ¿cómo ha sido el dolor de espalda en el peor de los casos?	10.40	0.8
A4. Durante los últimos siete días, ¿cómo ha sido el dolor de espalda en otros momentos?	4.31	0.4
A5. Durante los últimos siete días, ¿ha dormido mal a causa del dolor de espalda?	3.66	0.8

Tabla 4.- Valores y puntos de máxima información para la dimensión Dolor

El ítem **A3** (*Durante los últimos siete días, ¿cómo ha sido el dolor de espalda en el peor de los casos?*) que también aporta información la hace en las categorías intermedias y altas y el ítem **A2** (*Durante los últimos siete días, ¿cuánto tiempo le ha durado el dolor de espalda durante el día?*) aunque aporta información en las categorías centrales no es una curva “normal” (Fig. 3).

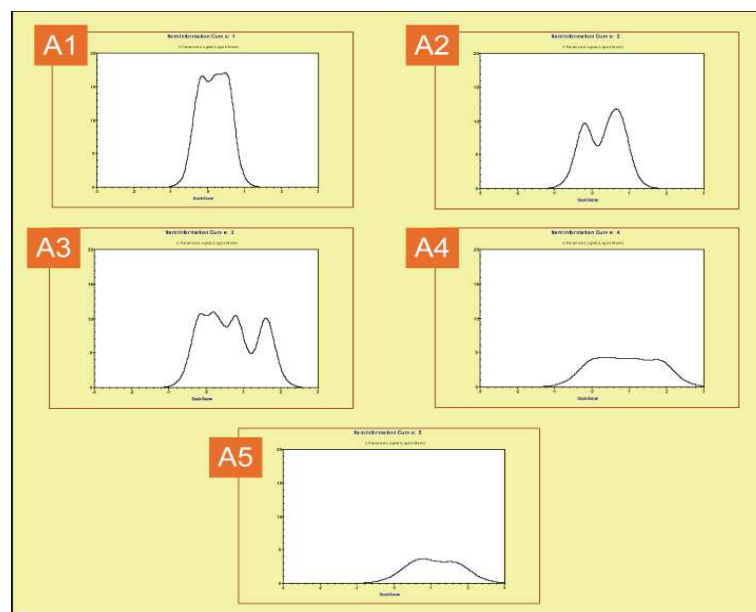


Figura 3.- Función de Información de cada uno de los ítems de la dimensión Dolor. Tomada de SANCHEZ_BARBA, 2008

Estimación de los parámetros y curvas características

La Tabla 5 muestra la estimación del parámetro de discriminación y los parámetros de dificultad para cada uno de los ítems de la dimensión del *Dolor*.

DOLOR					
ÍTEMS	Parámetro a	Parámetro b_1	Parámetro b_2	Parámetro b_3	Parámetro b_4
A1	7.82	-0.22	0.14	0.48	0.55
A2	6.17	-0.20	0.50	0.70	0.78
A3	6.34	-0.18	0.25	0.80	1.61
A4	3.85	0.00	0.53	1.15	1.83
A5	3.45	0.57	0.70	0.92	1.66

Tabla 5.- Valores de los parámetros de los ítems de la dimensión Dolor

Los ítems **A1** (*Durante los últimos siete días, ¿con qué frecuencia ha tenido dolor de espalda?*), **A3** (*Durante los últimos siete días, ¿cómo ha sido el dolor de espalda durante el día?*) y **A2** (*Durante los últimos siete días, ¿cuánto tiempo le ha durado el dolor de espalda durante el día?*) son los que mayor poder tienen para discriminar (parámetros de discriminación más altos) entre pacientes con distintos grados de Calidad de Vida. Estos ítems coinciden con los que más información aportan a la dimensión del *Dolor*. No obstante la interpretación de los parámetros de dificultad pone de manifiesto que esto es cierto para pacientes que evalúan su Calidad de Vida en intervalos que van de -0.22 a 0.55, para el ítem **A1**, de -0.18 a 1.61 para el **A3** y de -0.20 a 0.78 para el ítem **A2**. Esto significa, por ejemplo que el ítem **A1** aporta información para los pacientes que tienen una Calidad de Vida media (satisfactoria o regular en la escala presentada anteriormente) y no aporta información para los pacientes que tienen una buena (o muy buena) Calidad de Vida, ni tampoco para los que tienen una mala (o muy mala) Calidad de Vida. Ver más detalles en SANCHEZ-BARBA, 2008

Los ítems **A4** (*Durante los últimos siete días, ¿cómo ha sido el dolor de espalda en otros momentos?*) y **A5** (*Durante los últimos siete días, ¿Ha dormido mal a causa del dolor de espalda?*) tienen menor poder discriminativo ya que sus parámetros de discriminación son considerablemente más bajos. Esto puede ser debido a que en el ítem **A5** se mezcla el concepto del *Dolor* con el concepto del *Sueño* y el ítem **A4** puede ser para nuestros pacientes algo confuso, se les pregunta por el dolor que han padecido en otros momentos que el que han padecido en el peor de los casos y para ellos lo importante es el momento de máximo dolor, recordemos que más de 65% de los individuos tienen más de 60 años.

En el siguiente gráfico vienen representadas las curvas características de los cinco ítems que componen la dimensión del *Dolor*.

El ítem **A1** (*Durante los últimos siete días, ¿con qué frecuencia ha tenido dolor de espalda?*) es el que aporta más información, sin embargo, no todas las categorías de respuesta del ítem son igualmente portadoras de información; concretamente la categoría 4 (*4-6 días*) carece de interés porque para ninguno de los niveles de Calidad de Vida del individuo esta categoría presenta una alta probabilidad de ser elegida. El

ítem **A2** (*Durante los últimos siete días, ¿cuánto tiempo le ha durado el dolor de espalda durante el día?*) también aporta información relevante (aunque menor que la aportada por el ítem **A1**), pero en este caso dos de las categorías (la 3 y 4 de 3 a 5 horas durante el día y de 6 a 10 durante el día) podrían ser suprimidas por la baja probabilidad de ser elegidas.

El ítem **A3** (*Durante los últimos siete días, ¿cómo ha sido el dolor de espalda en el peor de los casos?*) que también aporta información presenta la ventaja de que todas las categorías están bien diferenciadas. Los ítems **A4** (*Durante los últimos siete días, ¿cómo ha sido el dolor de espalda en otros momentos?*) y **A5** (*Durante los últimos siete días, ¿ha dormido mal a causa del dolor de espalda?*) cuyas categorías de respuesta están bien diferenciadas, presentan información irrelevante a la hora de autoevaluar la Calidad de Vida percibida.

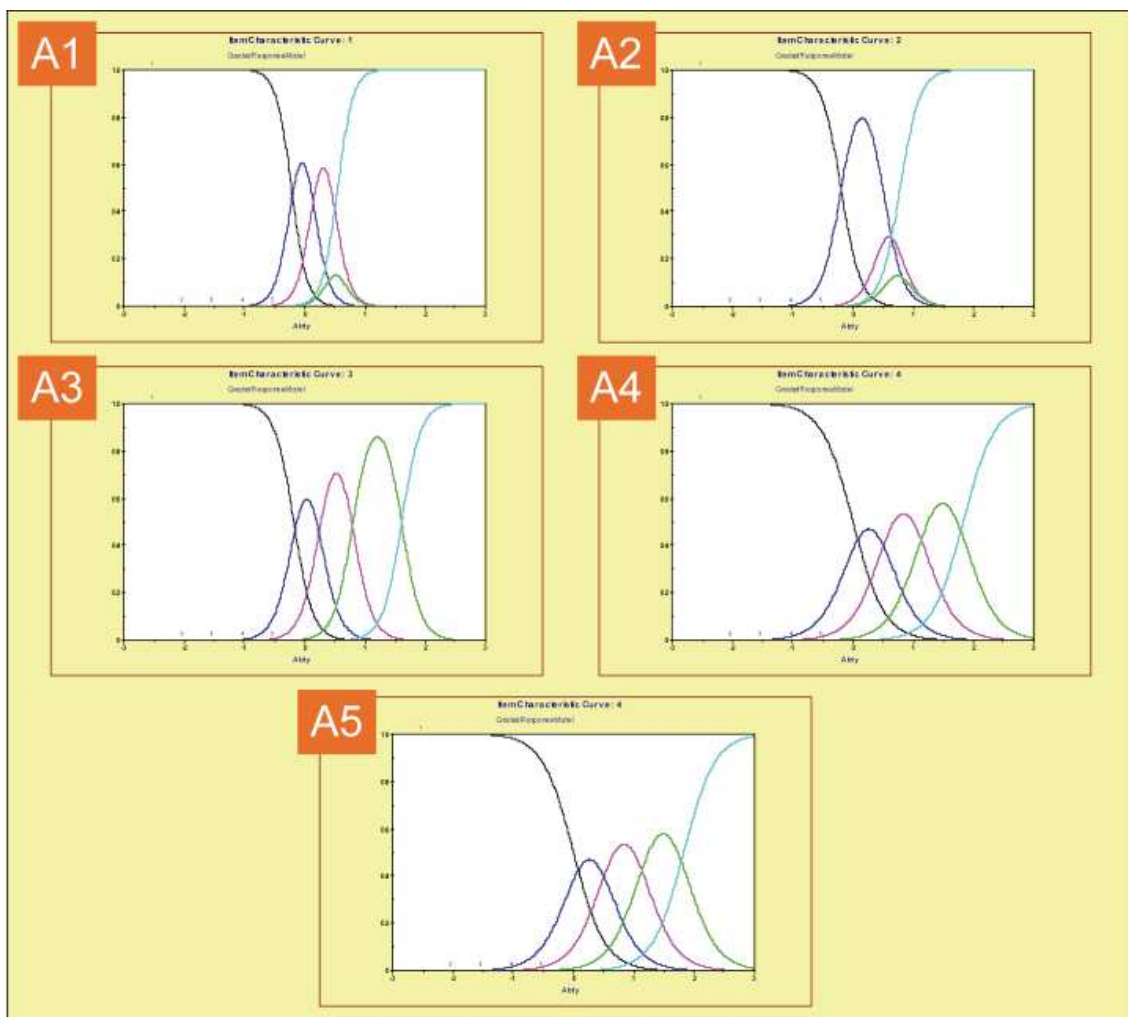


Figura4: Curvas características de los ítems de la dimensión Dolor. Tomada de SANCHEZ-BARBA 2008

3.4. VALIDEZ FACTORIAL *versus* IMPACTO *versus* CAPACIDAD DISCRIMINANTE

3.4.1. Dimensión Dolor Del Qualeffo

Comparando los resultados encontrados en el estudio de la Validez Factorial, el Impacto y Capacidad discriminante de los ítems de la dimensión Dolor, del cuestionario QUALEFFO, encontramos:

Todos los ítems de la Dimensión Dolor conforman una misma dimensión latente (eje factorial 2); el ítem con mayor carga factorial es el 3, pero el A1, el A2 y el A4 tienen cargas similares a la del A3. La carga del ítem A5 es considerablemente más baja. Los ítems A1, A2 son los que presentan máxima frecuencia, máxima importancia y máximo impacto. Las características del ítem A3 son similares a las de los anteriores, pero la importancia y el impacto del A4 y el A5 son considerablemente más bajas. Los ítems A1, A3 y A2 son los que mayor poder tienen para discriminar (parámetros de discriminación más altos) entre pacientes con distintos grados de Calidad de Vida. Estos ítems coinciden con los que más información aportan a la dimensión del *Dolor* (según la TRI). Luego los ítems A1, A2 y A3 son buenos candidatos a ítems que deben figurar en el cuestionario simplificado. Si queremos quedarnos con dos tal como sugiere ALLEN Y LOCKER, 2002, habría que decidir si quedarse con el A2 o el A3. Si primamos el impacto nos quedaríamos con el A1 y el A2, pero si tenemos en cuenta que en el ítem A2, las categorías 3 y 4 tienen una bajísima probabilidad de ser elegidas y que en el ítem A3 todas las categorías aportan información, probablemente fuese más interesante mantener el ítem A3 en el cuestionario más breve. Lo que si es evidente es que la información con ambos métodos no es equivalente sino más bien complementaria.

Análogamente podríamos estudiar las demás dimensiones. El estudio detallado de la validez factorial y del impacto se incluye en este trabajo. Los resultados con los modelos de la TRI pueden consultarse en SANCHEZ BARBA, 2008.

3.4.2. Dimensión actividades cotidianas, del Qualeffo

El estudio comparativo de los tres métodos nos permite afirmar lo siguiente:

La dimensión B no está bien definida según el estudio de la estructura latente, B6 y B7 cargan alto en dos ejes y el B8 carga en un eje diferente. El B8 es el ítem que tiene mayor frecuencia, es el segundo con mayor importancia y es el que tiene más impacto, seguido, a distancia del B7. Por tanto desde el punto de vista de la consistencia factorial se podrían seleccionar el B6 y el B7 que tienen una estructura más homogénea, pero desde el punto de vista del impacto deberíamos elegir el B7 y el B8. Aplicando los modelos de la TRI, el ítem B6 (¿Tiene dificultad para vestirse?) es el que aporta más información (casi el doble de la media esperada), y lo hace en las categorías más altas. El ítem B8 (¿Tiene dificultad para dormirse a causa del dolor?) no aporta información relevante a la dimensión, (la función de información coincide prácticamente con el eje de abscisas), por lo que es un claro candidato a ser eliminado al cuestionario QUALEFFO. El ítem B6 (¿Tiene dificultad para vestirse?) es el que más discrimina entre los pacientes con distintos grados de Calidad de Vida, aunque según ponen de

manifiesto sus parámetros de dificultad lo hace en los pacientes que tienen una mala (o muy mala) Calidad de Vida. El ítem B7 (¿Tienen dificultad para bañarse o ducharse?) también discrimina entre pacientes con una mala Calidad de Vida aunque aporta menos información para valorar ese nivel de calidad de Vida que el ítem B6. El ítem B8 (¿Tiene dificultad para dormirse a causa del dolor?) tiene poco poder para discriminar.

3.4.3. Dimensión tareas domesticas, del Qualeffo

El estudio comparativo de los tres métodos nos permite afirmar lo siguiente:

Desde el punto de vista de la estructura factorial, todos los ítems cargan en un mismo eje, pero no solo ellos. Desde el Análisis del Impacto concluimos que los ítems con mayor impacto son el C12 y el C13. El ítem C12 (¿Puede hacer la compra diaria?) no es el que aporta más información (desde la TRI) aunque si tiene interés en las categorías intermedias y altas. Sin embargo, el ítem C13 (¿Puede levantar un objeto pesado de 10Kg y trasladarlo por lo menos 10 metros?) no aporta información relevante a la dimensión Tareas Domésticas. Los ítems C10 (¿Puede hacer la comida?) y C11 (¿Puede lavar los platos?) son los que más información aportan a la escala Tareas Domésticas, superior a la media de información que se esperaría para cada ítem, según el estudio de SANCHEZ-BARBA, (2008) basado en la TRI. De nuevo encontramos discrepancias importantes entre ambos métodos.

3.4.4. Dimensión movilidad, del Qualeffo

El estudio comparativo de los tres métodos nos permiten afirmar lo siguiente: Desde el punto de vista de la estructura factorial, todos los ítems cargan en un mismo eje, pero no solo ellos. Desde el Análisis del Impacto concluimos que los ítems con mayor impacto son el D15 y el D16.

Para esta dimensión, la información aportada por los ítems D14 (¿Puede levantarse de la silla?) y D17 (¿Puede andar 100 metros?) es prácticamente la misma y está en torno a la media de información esperada que debería de corresponder a cada ítem. La información aportada por el ítem D15 (¿Puede inclinarse?) es inferior.

La información que aportan estos tres ítems lo hacen en las categorías centrales y altas aunque no son curvas “normales”. El ítem que menos información aporta a la dimensión Movilidad es el ítem D19 (¿En qué medida le han afectado los cambios en su aspecto físico debidos a la osteoporosis?), probablemente se debe al hecho de que el cuestionario está pasado a sujetos que acuden a una consulta de Atención Primaria y el estudio está diseñado con fines exploratorios, por lo tanto los pacientes no están realmente afectados por estos cambios y por eso no son importantes en su percepción de Calidad de Vida.

3.4.5. Dimensión actividades sociales y tiempo libre, del Qualeffo

El estudio comparativo de los tres métodos nos permite afirmar lo siguiente:

Desde el punto de vista de la estructura factorial, todos los ítems cargan en un mismo eje, pero no solo ellos. Desde el Análisis del Impacto concluimos que los ítems con mayor impacto son el E20 y el E21.

Desde los modelos de la TRI se deduce que el ítem que más información proporciona es el E22 (¿Puede visitar a familiares y amigos?), y el que menos el E21 (¿Puede ir al cine, al teatro, a un restaurante...?), diez puntos menos de la que le debería de corresponder en media. La información aportada por el ítem E23 (¿Puede llevar a cabo actividades sociales?) está en torno a la media esperada para cada ítem.

De nuevo nos encontramos que el ítem más importante con un procedimiento es el menos importante con el otro.

3.4.6. Dimensión estado de ánimo, del Qualeffo

El estudio comparativo de los tres métodos nos permite afirmar lo siguiente:

Desde el punto de vista de la estructura factorial no se encuentra la estructura esperada; más bien parece que los ítems de esta escala están conformando dos dimensiones latentes.

Desde el Análisis del Impacto concluimos que los ítems con mayor impacto son el G29 y el G30.

Desde la TRI, los ítems que más información aportan a la dimensión Estado de Ánimo, están relacionados con el Miedo a las Fracturas. El ítem que más información aporta es el G34 (Durante los últimos siete días, ¿Ha tenido miedo a caerse?), seguido del G33 (Durante los últimos siete días, ¿Ha tenido miedo de tener una fractura?) y G35 (Durante los últimos siete días, ¿Ha disminuido su actividad por miedo a las fracturas?). Esta información está por encima de la media esperada. El ítem G30 (Durante los últimos siete días, ¿Ha visto el futuro con esperanza?) es el que menos información aporta y además destacar que el valor de máxima información no lo alcanza en un punto sino a lo largo de un intervalo.

En la Figura 5 vienen representadas las funciones de información de los ítems G29 (Durante los últimos siete días, Se ha sentido animado/a (con mucha energía)) y G30 (Durante los últimos siete días, ¿Ha visto el futuro con esperanza?), que no aportan información relevante a la dimensión Estado de Ánimo. Estos dos ítems son claros candidatos a ser eliminados del cuestionario QUALEFFO.

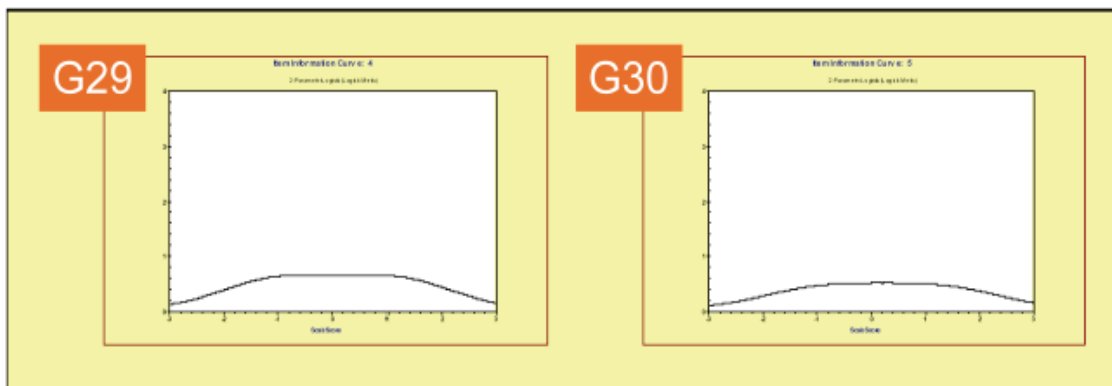


Figura 5: Función de información de los ítems G29 y G30. Tomada de SANCHEZ-BARBA 2008

3.4. Propuesta de versión simplificada del Qualeffo

De todo el estudio anterior podemos afirmar una vez más, que los ítems con mayor impacto no tienen capacidad informativa desde el punto de vista de la discriminación. La tabla 6 recoge todos los resultados de los apartados anteriores y a partir de la información de esta tabla, hacemos nuestra propuesta de simplificación del Cuestionario QUALEFFO para Atención Primaria.

	<i>Con el AFE</i>	<i>Impacto</i>	<i>TRI</i>
DIMENSIÓN			
<i>Dolor</i>	A1 y A3	A1 y A2	A1 y A2
<i>Actividades Cotidianas</i>	B6 y B7	B7 y B8	B6 y B7
<i>Tareas Domésticas</i>	C10 y C11	C12 y C13	C10 y C11
<i>Movilidad</i>	D14 y D18	D15 y D16	D14 y D17
<i>Actividades Sociales y T. Libre</i>	E22 y E23	E20 y E21	E22 y E23
<i>Salud General</i>	F24 y F25	F24 y F25	F24 y F25
<i>Estado de ánimo</i>	G33 y G34	G29 y G30	G33 y G34

Tabla 6.- Ítems que deberían formar el cuestionario QUALEFFO reducido, según los tres métodos estadísticos

Por todo ello nuestra propuesta para el QUALEFFO simplificado, adaptado a pacientes de Atención Primaria, constaría de los siguientes ítems: Dolor (A1, A2 y A3); Actividades Cotidianas (B6, B7 y B8); Tareas domésticas (C10, C11, C12 y C13); Movilidad (D14, D15, D16, D17 y D18); Actividades sociales y Tiempo Libre (E20, E21, E22 y E23); Salud General (F24 y F25) y Estado de Animo (G29, G30; G33 y G34), la cual tiene alta validez factorial, alta validez de contenido y alto poder discriminante.

Esta propuesta resuelve, además, otro de los principales problemas al ser aplicado en Atención Primaria, el tiempo necesario para su aplicación, ya que el cuestionario original tenía 35 ítems y la versión simplificada propuesta solo 225.

CONCLUSIONES

La revisión de la literatura pone de manifiesto que el método más utilizado para la propuesta de cuestionarios reducidos es el Análisis Factorial Exploratorio. El análisis del impacto se ha utilizado en Calidad de Vida relacionada con la Salud Dental presentándolo como un método alternativo mejor que el clásico Análisis Factorial, para la simplificación de cuestionarios.

La aplicación de la Teoría de Respuesta al Ítem es menos utilizado que el Análisis Factorial, pero mucho más que el Análisis del Impacto de los Ítems.

El análisis Factorial aplicado al QUALEFFO, en pacientes revisados en Atención Primaria, no proporciona resultados similares a los encontrados en pacientes evaluados en el contexto hospitalario.

Los ítems con mayor impacto no son los ítems con mayor carga factorial, lo cual era esperable ya que los ítems con más impacto son generalmente los más frecuentes y para esos ítems no hay mucha variabilidad, siendo esta la esencia del Análisis Factorial.

Los ítems seleccionados con el Análisis Factorial son los ítems más relevantes, si estamos interesados en la validez factorial, y los ítems con más impacto los más asociados con la Validez de Contenido. Los ítems seleccionados con la TRI no son los de más impacto, lo cual era esperable ya que si el síntoma asociado al ítem lo presentan la mayoría de los pacientes, ese ítem no es bueno para discriminar entre pacientes con distintos niveles de Calidad de Vida. Los tres procedimientos, que hasta ahora se han usado por separado, deberían utilizarse conjuntamente para poder disponer de cuestionarios con alta validez factorial, alta validez de contenido y alto poder discriminante.

BIBLIOGRAFIA

ABIZANDA, M. (2005). Recomendaciones generales sobre la osteoporosis. Manejo de la Osteoporosis en Atención Primaria. Sociedad Española de Medicina Rural y Generalista y Scientific Communication Management, S.L.

AHMAVAARA, Y. (1954). Transformation analysis of factorial data. *Annals of the Academy of Science Fennicae*. Series B. 881 (2), pp: 54-59.

ALLEN, F.; LOCKER, D. (2002). A Modified Short Version of the Oral Health Impact Profile for Assessing Health-Related Quality of Life in Edentulous Adults. *The International Journal of Prosthodontics*; 15(5). pp: 446-450.

BADÍA, X.; HERDMAN, M. (1999). Adaptación transcultural al español de los cuestionarios OQLQ y QUALEFFO para la evaluación de la calidad de vida relacionada con la salud de mujeres con fractura vertebral osteoporótica. *Rev Esp Enf Met Oseas*. 8. pp: 135-140.

- BIRNBAUM, A. (1968), Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. pp: 397-472. (In LORD, F.M., NOVICK, M.R. *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley).
- COHEN, A.S.; KIM, S.H.; BAKER, E. (1993). Detection of Differential Item Functioning in the Graded Response Model. *Applied Psychological Measurement*. 17. pp: 335-350.
- JUNIPER, E.F.; GUYATT, GH. STREINER, DL; KING, D.R. (1997). Clinical impact versus factor analysis for quality of life questionnaire construction. *J Clin Epidemiol*. 50. pp: 233-238.
- LIPS, P.; COOPER, C.; AGNUS, D.; CAULIN, F.; JOHNNELL, O.; KANIS, J. y col. (1999). Quality of life with vertebral fractures: validation of the Quality of Life Questionnaire of the European Foundation for Osteoporosis (QUALEFFO). *Osteoporosis Int*. 10. pp: 150-60.
- MURRELL, P.; TODD, C.J.; MARTIN, A.; WALTON, J.; LIPS, P.; REEVE, J. (2001). Postal administration compared with nurse-supported administration of the QUALEFFO-41 in a population sample: comparison of results and assessment of psychometric properties. *Osteoporosis Int*. 12(8) pp: 672-9.
- SAMEJIMA, F. (1969). Calibration of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 17.
- SÁNCHEZ BARBA, M. (2008). *Aportaciones al análisis de datos de calidad de vida relacionada con la salud desde una perspectiva multivariante*. Tesis Doctoral. Universidad Salamanca
- STREINER, D.L.; NORMAN, G.R. (1995). *Health measurement scales: a practical guide to their development and use*. 2ed. Ed. Oxford University Press. Oxford.
- THISSEN, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory*. Ed. Scientific Software Internacional, Inc. Chicago.
- VICENTE-GALINDO, P. (2003). *Contribuciones al análisis de datos de Calidad de Vida Relacionada con la Salud*. Tesis Doctoral. Universidad de Salamanca.

Capítulo 3

REVISION DE ALGUNOS PROBLEMAS ESTADISTICOS EN LA ESTIMACION DE LA BIODIVERSIDAD

CARLOS N. BOUZA HERRERA* y DANTE COVARRUBIAS MELGAR**

*Universidad de La Habana, Cuba

**Universidad Autónoma de Guerrero, México

Resumen

Se hace una presentación de los modelos que son de uso común en el estudio de la biodiversidad. Se hace hincapié en el problema que parece al hacer la estimación usando muestras aleatorias. Se intenta que este trabajo sirva de referencia para posteriores desarrollos de modelos particulares para los estudios de la biodiversidad.

Palabras clave: estimación, errores de muestreo, índices de biodiversidad, índices de cubrimiento, rareza.

Abstract

We present models of common use in biodiversity studies. We highlight the problems present in estimating using random samples. We intend that this paper will play the role of being a reference for posterior developments of particular models for studies of biodiversity.

Keywords: estimation, sampling errors, indexes of biodiversity, indexes of coverage, rarity.

1. ENFOQUES Y CRITERIOS PARA LA PLANEACION Y DESARROLLO DEL LEVANTAMIENTO DE LOS DATOS

1.1. Enfoques para la experimentación

En el estudio de la biodiversidad podemos usar diversos enfoques. Ellos se asocian a varios requerimientos básicos a tomar en cuenta al planear la recolección de datos. Estos son detallados en la tabla siguiente:

Inferencia basada en el diseño	Inferencia basada en el Modelo
Se requiere tamaños de muestra relativamente grandes	Se puede usar tamaños de muestra pequeños
Se tiene que seleccionar muestras aleatorias	Se pueden seleccionar muestras representativas usando el modelo o un experto
La población objetivo está bien determinada	Se hacen consideraciones sobre clases de problemas, Fuentes o tendencias de cambios
Se puede hacer comparaciones entre poblaciones	Se hacen inferencias sobre relaciones bajo el modelo o se determinan modelos alternativos
Puede ser costosa para mantener niveles de precisión prefijados	Puede ser cara en términos de las dificultades de hallar sitios representativos.

Tabla 1.1 Enfoques del proceso de toma de datos

Podemos caracterizar los enfoques y construir la tabla 2.2. El uso del principio inferencial basado en el diseño es adecuado al tomar en cuenta el uso de la teoría del muestreo lo que permite tipificar la toma de datos. El procedimiento inferencial es soportado por la teoría del muestreo repetido que permite utilizar o desarrollar nuevos métodos de la llamada estadística clásica. Al tener tamaños de muestra suficientemente grandes se puede garantizar la validez del uso de la teoría normal y en todo caso algún método no paramétrico. Esto permite hacer comparaciones entre las medidas de biodiversidad de varias poblaciones. La posible necesidad de tener tamaños de muestra grandes para aplicar modelos normales encarece su uso.

Por otra parte el uso de modelos para hacer la inferencia permite escapar de la necesidad de tomar muestras aleatorias. La representatividad viene dada por criterios asociados al modelo asumido como generador de las variables o de la medida de biodiversidad. En las aplicaciones de biodiversidad se denominan tales sitios como 'centinelas' Estos son unos pocos y en ocasiones son utilizados para construir un modelo. Este es el caso al monitorear. Cada sitio representa un problema o clase de problemas de interés para la biodiversidad y los estudios desarrollados en ellos son intensivos. Generalmente son usados factores ecológicos, tales como la sensibilidad a ciertas tensiones o tendencias. De tales estudios se genera un modelo que se considera válido para otros sitios y se usa para hacer la inferencia. Su debilidad es que buscar los sitios representativos puede ser muy difícil. Los centinelas pueden estar en regiones de difícil acceso y su identificación ya es de por sí complicada. En todos los casos hacer generalizaciones de experiencias basadas en modelos es muy riesgosa dada la variabilidad de la biodiversidad en el mundo real.

Enfoque Operacional	Enfoque Experimental
Manejo Pasivo y Adaptativo	Manejo Activo y Adaptativo
Usa sitios disponibles y operacionales	Crea sitios y tratamientos
Hace comparaciones entre los resultados de la operación común en curso (Cosecha, pesca, tala etc.,)	Hace comparaciones entre posibles formas de operar
Limita el número de tratamientos a los que posibilita el manejo	Fija un número de tratamientos y se adapta el manejo a lo requerido por ellos
Es relativamente barato	Puede ser costoso
No hace propuestas con efecto profundo en el manejo	Hace propuestas que pueden conllevar profundos cambios en el manejo

Tabla 1.2 Caracterización de enfoques de los experimentos

El enfoque operacional es de carácter pasivo- adaptativo y maneja sitios operacionales de disponibilidad inmediata. Es útil en el caso de buscarse hacer comparaciones en cosechas, problemas forestales en lo que es importante el monitorear el manejo. Este incluye estudios retrospectivos para conocer la historia de la biodiversidad a través de los datos acumulados. Por ello los métodos de monitoreo son muy usados, lo que conlleva que no sean útiles para comparar tratamientos. Estos llevan a cabo recomendaciones útiles para el manejo adaptativo de los sistemas.

Los estudios de biodiversidad buscan aprender o comprender los mecanismos que operan sobre ella en el hábitat escogido. Los datos por si mismos no dan conocimiento de este tipo. Al desear manejar el sistema se buscará explicar el mecanismo que le rige para poder operar sobre este. Para ello se pueden hacer comparaciones para aprender cómo opera el mecanismo o simplemente explicar el funcionamiento del mismo. Podemos decir que uno es de carácter indagatorio y el otro de carácter mecanicista.

1.2 Diseño estadístico de redes de monitoreo

El objetivo del monitoreo es detectar los cambios en los parámetros de interés en un periodo relativamente largo. Obtener una guía aceptable se vincula con la inferencia estadística. Adicionalmente es usual que se haga el diseño del estudio de monitoreo sin consultar a un estadístico lo que agrava aún más las posibilidades reales de hacer inferencias confiables de los resultados. Claramente en ecología los experimentos y el monitoreo plantean problemas conceptuales a la Estadística que no siempre están resueltos satisfactoriamente. Por ejemplo diseñar una investigación en gran escala, las variaciones no controladas presentes (temperatura, incendios etc.). Dificultad de obtener réplicas, grandes variaciones los modelos de muestreo razonablemente. Por este motivo los programas de monitoreo están, generalmente, deficientemente diseñados desde el

punto de vista estadístico y el método de medición es generalmente poco confiable. Es difícil hacer un diseño de monitoreo que satisfaga los requerimientos básicos para hacer inferencias confiables.

El ecólogo está generalmente interesado en que el estadístico le diseñe un programa que sea robusto desde el punto de vista inferencial. Esto es soportado por hacer un diseño tal que:

- Compare las diversas posibilidades de muestrear usando diversos tratamientos.
- Garantice un número suficiente de réplicas para tratamientos tomando en cuenta la heterogeneidad de los ecosistemas y la incertidumbre.-Haga una asignación usando criterios de estratificación para garantizar que haya alguna homogeneidad entre las unidades de las subzonas.
- Tome información suficiente (tiempo de monitoreo alto) para poder establecer el efecto de los tratamientos aislándolas causas externas como cambios climáticos.

2. MEDIDAS DE BIODIVERSIDAD

2.1 Generalidades

Existen diversas medidas para caracterizar la biodiversidad. Muchas de ellas no tienen un tratamiento estadístico debido al hecho mismo de que fueron propuestas sin tomar en cuenta el uso de estudios estadísticos, pensando en hacer inferencias. Los ecólogos usan medidas para caracterizar la biodiversidad y evaluar la salud de los ecosistemas. Su disminución indica la del ecosistema y llama la atención sobre el stress medio ambiental del mismo. La diversidad es evaluado mediante un sin número de medidas, propuestas puntualmente por ecólogos.

La pérdida en diversidad biológica por extinción es uno de los cambios más notables en la transformación global del medio ambiente hoy día. Los investigadores tratan de comprender el mecanismo de los cambio en la biodiversidad.

Una visión hacia las políticas actuales de los países permite medir el nivel de importancia que a cobrado la ecología en la actualidad. El creciente interés por el tema en el mundo y el apoyo de diferentes organismos internacionales manifiestan una verdadera preocupación por conocer y preservar nuestros recursos naturales

Al medir diversidad es necesario tomar en cuenta que se trata de describir la relación entre individuos de diferentes especies en una zona. El número de ellos es medido. Algunas asunciones deben ser hechas al considerar le evaluación de la biodiversidad

Las categorías son conocidas: La clasificación de las especies es de universal aceptación-

Todas las categorías son igualmente diferentes: No hay error posible de clasificar un individuo en más de una clase.

Usar una medida de la importancia de las especies: Usualmente es una medida que usa el cubrimiento y número de individuos, densidad relativa.

La comunidad bajo estudio está bien definida: Como la importancia relativa de una categoría variara dependiente de la definición de su tamaño y extensión de la zona.

De acuerdo con (Smith, 2002) a lo largo de las últimas cuatro décadas se han registrado avances importantes en los estudios ecológicos y particularmente de estadística en la ecología, ejemplos de ello son las contribuciones al estudio de las distribuciones espaciales de organismos y los diseños de estudios ecológicos.

La implementación de políticas que incorporan caracterizaciones del estado que guarda una comunidad ecológica, hacen uso de indicadores propuestos varias décadas antes, los *índices de riqueza, equilibrio y diversidad*.

El concepto de diversidad ha sido ampliamente discutido como se puede ver en la literatura especializada. Un aporte a esta discusión la hacen Patil y Talle (1982) quienes consideran a la diversidad como una propiedad medible intrínseca de la comunidad, definida como el promedio de la rareza de especie.

2.2 Algunos índices

Los índices de biodiversidad más frecuentemente utilizados utilizan medidas de riquezas o rareza y la distribución de las especies. Ejemplos de ellos son los de Simpson, (1949), Shannon (1948), Brillouin (1956), McIntosh (1967), Hurlbert (1971) y Fager (1972). Estos son re-escalados en diferentes trabajos por los ecólogos. Tendrá sentido hacerlo para garantizar propiedades estadísticas pero este no es el motivo usualmente. Veremos algunas de ellas

2.2.1 La riqueza

La riqueza de las especies es la medida más simple. Es intuitivo considerar que este es la forma correcta de establecer la diversidad objetivamente. Esta es usada comúnmente en estudios conservacionistas para determinar la sensibilidad de ecosistemas en las especies que le habitan. Es usada como una primera medida de la diversidad en el hábitat. Fue llamado así por McIntosh (1967). El conteo puede ser complicado en muchas ocasiones. En general es útil al comparar dos hábitats. Ninguna inferencia puede ser hecha usándole.

2.2.2 Rareza

Esta es medida por la razón entre el valor de un índice de diversidad y su valor máximo. Este mide como se distribuye la abundancia entre las especies de una comunidad. El más usado es el referido al índice Shannon en el que se fija el parámetro como

$$E(\text{Shannon}) = \frac{-\sum_{i=1}^K \pi_i \ln(\pi_i)}{\ln(K)}$$

Por su definición, $E \in [0,1]$. Este es conocido muchas veces como índice de Pielou.

2.2.3 El índice de Brillouin

Para muestras no aleatorias es muy usado y es preferido a al índice de Shannon. Bajo el supuesto de que K es conocido sin error, este es

$$B = \frac{\ln(N!) - \ln(N_i!)}{N} B$$

Este fue usado por vez primera por Margalef (1958).

2.2.4 Métodos del tipo Monte Carlo

Se sabe que generalmente hay una correlación inversa entre riqueza y latitud: así como entre esta y la altitud. La riqueza de las especies no se relaciona con el tamaño del área estudiada y la observación de más especies puede deberse solo al uso de muestras más grandes. Áreas mayores por su parte comprenden un mayor número de hábitats dando más oportunidad a más especies a asentarse debido a la mayor diversidad. Esta medida es utilizada con frecuencia en estudios taxonómicos de comunidades o grupos funcionales. Tome como ejemplos la flora de un bosque (taxonómico) y la del zooplancton (funcional). Los principios del Jackknife y del Bootstrap son utilizados para examinar el proceso que refleja. Estos tratan de establecer el número de especies faltantes en la muestra. Para ello se evalúa en cada re-muestra el número de especies faltantes para estimar cuantas serian no detectadas.

El ajuste de una ecuación que describa la acumulación de las especies permite graficar una curva, la que se utiliza para predecir el número de especies. Los expertos utilizan a veces inopinadamente modelos como la ecuación de Michaelis-Menton. Esta describe la reacción de las enzimas-catalizadas y no es claro que sirva igualmente para describir la riqueza de animales o plantas, aunque se haga con frecuencia. Tiene más sentido hacer un ploteo con los datos obtenidos y alizar la curva usando algún método numérico. Un ejemplo son los trabajos de Rennolls- Laumonier (2006), quienes trabajaron con especies de árboles en Sumatra. Ellos asumieron que las especies raras en los sitios eran una muestra aleatoria de estas en la selva y estimaron el número de especies ocultas.

Un métodos novedoso es el aplicado por Royle-Dorazio (2008) quienes usaron los criterios Bayesianos implementados en el software comercial para fijar la probabilidad de detectar las especies y con ello estimar el total de ellas. La acumulación de especies permite juzgar como los estimadores de la riqueza trabajan para nuestros datos. La línea recta es el resultado estimado, en el que las especies se equilibran.

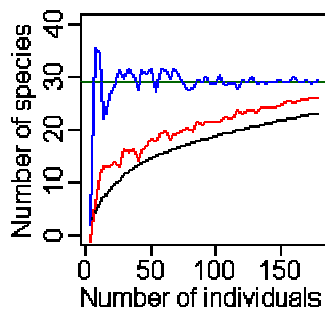


Figura 2.1. Ejemplo de un gráfico de la riqueza

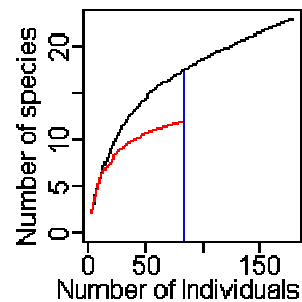


Figura Ejemplo de un gráfico de curvas alisadas

Las líneas azul y roja describen como se comporta la riqueza al aumentar el número de individuos observados en dos estimadores. Su comparación permite valorar las

diferencias entre dos zonas estudiadas. La azul refleja un comportamiento muy bueno. La negra describe el proceso teórico de nivelación de la riqueza.

En general es necesario responder si hay más especies y cuantas son. Para ello será necesario inferir por interpolación. Por ejemplo si en un sitio observamos 80 insectos de 10 y en otro 174 de 22 se desea establecer si hay más especies en el primer sitio. Tomamos k submuestras $s(i)$ del segundo sitio de tamaño 80 y se mide el número de especies observadas, $n(i)$.

$$\bar{n} = \sum_{i=1}^k \frac{n(i)}{k}$$

Este método es conocido como rarefacción (rarefaction), ver Kemton (2002).

En la figura siguiente tenemos las curvas alisadas de en los dos sitios. Podemos hacer la interpolación necesaria en la figura 3.2, usando este gráfico.. Es confiable utilizar la extrapolación solo si se tienen tamaño muestral grande.

2.2.5 Estimación de la cobertura

El problema de estimar o predecir el número de clases en que se divide una población U está presente en muchas aplicaciones. Lo usual en la Teoría de Muestreo es que el decisor conoce las características que determinan las subpoblaciones y su cantidad. Este es el caso de la estratificación, la conglomeración y los dominios de estudio por ejemplo. Consideremos que U está formada por un número desconocido Δ de clases. Se toma una muestra $s \subset U$ usando muestreo simple aleatorio (MSA). Los objetos observados son clasificados en clases al ser identificados y considerando la similitud y diferenciación entre los grupos que determinan ellos. Este problema difiere del problema de la post estratificación sensiblemente. En la post estratificación conocemos los estratos y clasificamos los objetos después de tomar la muestra. Ahora lo que deseamos es conocer el número de clases. Podemos considerar que esta es una tarea de los métodos de aglomeración comunes en la paquetería estadística. Sin embargo el problema de fijar un valor Δ es de gran importancia en muchas aplicaciones. Si consideramos los problemas de la arqueología veremos que ese es el objetivo del investigador en ciertos momentos. Él analiza las piezas encontradas, determina grupos y piensa que el número de grupos formados es menor que los que existieron. En particular en los problemas medio ambientales es común una problemática similar pues se quiere conocer cuántos contaminantes están afectando el hábitat. Algunos ejemplos de problemas tipos son:

- Identificar el número de especies en una región o los contaminantes presentes en los desperdicios que una empresa bota en un vertedero natural.
- Determinar el léxico personal de un autor o las enfermedades causadas por el nivel de toxicidad en el aire de una región de salud.
- Fijar los defectos presentes en las unidades de una línea de productos de una fábrica o las malformaciones en la vegetación circundante a una central atómica eléctrica.
- Establecer los tipos de plaga que están presentes en un cultivo o en las fuentes de los polutantes observados en un río.

El trabajo pionero en esta temática se remonta al trabajo del fundador de la estadística, Fisher et. Al. (1943) quien lo desarrolló estudiando un problema de Ecología aplicada. Posteriormente ese ha estudiado problemas diversos. Son muy conocidos los de Efron-Thisted (1976) , quienes estudiaron el léxico de Shakespeare para establecer la autenticidad de ciertas obras, los de Engen (1978) ,Chao (1981) y Brained (1982) quienes continuaron la investigación de modelos para hacer una estimación de Δ en unos problemas biológicos. Otro enfoque es considerar la existencia de un proceso estocástico que explica la aparición de las clases en la muestra. Esto es más popular al considerar el arribo de especies a un punto de monitoreo, vea Chao-Bunge (2002) y Chao-Lee (1992).El problema teórico que este problema plantea varias interrogantes a la estadística. Muchas de ellas no están resueltas. Una línea es determinar cotas para Δ o estimarlas. Otra es estudiar el cubrimiento obtenido, por la muestra seleccionada, de las clases desconocidas a priori. En este trabajo se abordará este último problema. Los estimadores propuestos por Good (1953) y por Esty (1986) serán estudiados y se proponen alternativas. Tanto Good (1953) como Esty (1986) desarrollaron Teorema Centrales del Límite para determinar intervalos confidenciales basados en la aproximación normal. Sus hipótesis son algo fuerte al basarse en que la observación de las clases es descrita por un modelo de equiprobabilidad. En este trabajo proponemos el uso de un modelo superpoblacional del tipo regresión. Su aplicación en este caso conlleva un Modelo Lineal Generalizado como descriptor del comportamiento del cubrimiento. Esto no depende de la hipótesis de equiprobabilidad para garantizar la distribución normal asintótica del predictor.

Para analizar el comportamiento de las diversas alternativas para caracterizar el cubrimiento de la muestra utilizamos datos de dos investigaciones. Una son los datos obtenidos del estudio de la infestación de campos de caña por una plaga en –cuba. E toro es un estudio de biodiversidad forestal en México.

Para modelar este problema Bouza- Covarrubias () hicieron la partición de la población U como

$$U = \cup_{i=1}^A U_i, \quad U_i \cap U_j = \emptyset, \text{ si } i \neq j.$$

Al seleccionar una muestra s de tamaño n observamos D subpoblaciones. El modelo estocástico considera que hay Δ urnas y estas son ocupadas al azar observándose $D \leq \Delta$ ocupadas. Claramente, D es un estadígrafo suficiente para Δ . El cubrimiento de una muestra es definido por el parámetro $\theta = \sum I_t \pi_t$, donde $\pi_t = \text{Prob}(U_t \text{ aparezca en la muestra})$ y

$$I_t = \begin{cases} 1 & \text{si el número de unidades de } U_t \text{ en } s \text{ es mayor que } 0 \\ 0 & \text{en otro caso} \end{cases}$$

Note que θ es un valor relacionado con la muestra y no con la población, pues es el cubrimiento de s . Este problema aparece también en los estudios de biodiversidad en los que el índice está condicionado a la muestra observada.

Good (1953) propuso estimar θ mediante $\theta_1 = 1 - P_1$, $P_1 = n_1 / n$, siendo n_1 el número de clases observadas una vez. Esto parece sorprendente a primera vista pero es el índice de cubrimiento usado por Engen (1978), Starr (1979) y Chao (1981).

Otra solución fue propuesta por Esty (1986) al considerar que el número de clases observadas y la real se relaciona mediante la ecuación $D = \Delta^* [1 - \exp(-n/\Delta^*)]$ y usar como estimador $\hat{\theta}_2 = D/\Delta^*$.

La conocida convergencia de ambos estimadores a θ se resume en el siguiente teorema:

Teorema 1. (Engen, Starr, Esty). Si $\pi_t = \pi$ para toda clase U_t , $n \rightarrow \infty$, $n/\Delta \rightarrow \infty$, $n/\Delta \rightarrow \alpha$ para $\alpha \in]0, 1[$ entonces se cumplen

1. $n^{1/2} (\hat{\theta}_1 - \theta)$ tiene como distribución aproximada la normal $N\left(0, \frac{\alpha^2(1-\alpha)\ln(1-\alpha)}{\alpha + (1-\alpha)\ln(1-\alpha)}\right)$
2. $n^{1/2} (\hat{\theta}_2 - \theta)$ tiene como distribución aproximada la normal $N(0, (1-\alpha)(\alpha - \ln(1-\alpha)))$.

Norris-Metter (1993) estudiaron el caso en que se seleccionan muestras con el fin de estimar el cubrimiento. Los resultados observados en s generan las variables Bernoulli

$$Z_{it} = \begin{cases} 1 & \text{si la clase } t \text{ -ésima se observa en la selección } i \\ 0 & \text{en otro caso} \end{cases}$$

y podemos calcular $n_t = \sum_{i=1}^n Z_{it}$. Esta es un Binomial con parámetros n y π_t . Sea N_j el número de clases observadas j veces en la muestra. Analizando N_0 , el número de clases no observadas, su esperanza es

$$E(N_0) = \sum_{t=1}^{\Delta} P(n_t = 0) = \sum_{t=1}^{\Delta} (1 - \pi_t)^n \geq \sum_{t=1}^D (1 - \pi_t)^n$$

Norris-Metter (1993) desarrollaron un análisis Bayesiano para derivar cotas inferiores para $E(N_0)$. Nosotros tomaremos $p_t = n_t/n$, el que estima insesgadamente a π_t , podemos usar el predictor derivado de los resultados anteriores $N^*_0 = \sum_{t=1}^D (1 - p_t)^n$. Es claro que $D + N^*_0 \cong \Delta$ por lo que un predictor ingenuo, que subestima el número de clases es $\hat{\Delta} = D + N^*_0$

Este corrige el sesgo de D y tiene uno menor.

Partiendo de este resultado, si seguimos las ideas del procedimiento de Esty (1986), podemos utilizar

$$\theta_3 = \frac{D}{\hat{\Delta}_0} = \left(1 + \frac{\hat{\Delta}_0}{D}\right)^{-1}$$

para predecir el cubrimiento. Note que este predictor es una variante del estimador de Esty (1986). Además si $n \rightarrow \infty$, $(1 - \pi_t)^n \rightarrow \bar{\omega}_t$, $\bar{\omega} > 0$, para al menos un $t \in \{1, \dots, \Delta\}$, entonces θ_3 tiene una distribución aproximadamente normal con media $[\sum_{t=1}^{\Delta} (1 - \pi_t)^n]^{-1}$ y varianza $V_3 = V(D) + V(\hat{\Delta}_0) - \rho \sqrt{V(D)V(\hat{\Delta}_0)}$

donde ρ es el coeficiente de correlación entre D y $\hat{\Delta}$. Las hipótesis utilizadas en el Teorema 1 son muy fuertes en la mayor parte de las aplicaciones. En los ejemplos utilizados en la introducción la equiprobabilidad del fenómeno que genera la observación de las distintas clases es poco realista. Ciertos contaminantes y polutantes aparecen con menor frecuencia que otros, algunas plagas son más prolíferas que otras. Algunas dolencias respiratorias son más comunes ante el enrarecimiento del aire por una cierta emisión de gases, etc.

Es poco realista pensar que podremos acudir a los expertos para que nos afirmen que la equiprobabilidad es válida, lo que difícilmente ocurre, que fijen un valor de α y nos aseguren la convergencia de los parámetros envueltos para grandes valores de n . Es más sencillo o esperable que los expertos puedan dar ideas sobre la relación existente entre una variable que identifique la clase y la probabilidad de ser detectada. Un modelo superpoblacional suficientemente sencillo es tomar

$$Y_{it} = \begin{cases} 1 & \text{si la observación } i \text{ pertenece a la clase } t \\ 0 & \text{en otro caso} \end{cases}$$

y valorar si es aceptable que esta se relacione con π_t mediante el modelo usado por Pothoofet.al. (1992)

$$Y_{it} = \pi_t + \varepsilon_{it}, \text{ donde } E(Y_{it}) = \pi_t \text{ y}$$

$$E(\varepsilon_{it} \varepsilon_{it'}) = \begin{cases} \sigma_t^2 & \text{si } Y_{it} = 1 \\ 0 & \text{en otro caso} \end{cases}$$

Ahora tenemos la variable Bernoulli

$$I(i, t|s) = \begin{cases} 1 & \text{si } Y_{it} = 1 \\ 0 & \text{en otro caso} \end{cases}$$

y $\hat{\pi}_t = \frac{\sum_{i=1}^n Y_{it}}{n}$ entonces tenemos que es modelo-insesgado el predictor $\theta_t = \frac{\sum_{t=1}^D \sum_{i=1}^n Y_{it}}{n}$ pues

$E(\hat{\pi}_t) = \pi_t$ y dada la independencia de las observaciones $V(\theta_t) = \sum_{t=1}^D \pi_t (1 - \pi_t) / n$. Note que el nuevo estimador es una función lineal de variables Binomiales independientes. Siguiendo las ideas de Bouza (1996) bajo el modelo superpoblacional se tiene que si $n \rightarrow \infty$ entonces $n^{1/2} (\theta_t - \pi_t) \rightarrow N(0, \sum_{t=1}^D \pi_t (1 - \pi_t) / n)$

Note que si $\pi_t = \pi$ para todo t la eficiencia de este predictor comparado con sus contrapartes y

$$E_{14} = \frac{-\alpha^2 (1 - \alpha) \ln(1 - \alpha) n}{\alpha + (1 - \alpha) \ln(1 - \alpha) D \pi (1 - \pi)}$$

y

$$E_{24} = \frac{(1 - \alpha)(\alpha - \ln(1 - \alpha)) n}{D \pi (1 - \pi)}$$

Ambas razones son mayores que uno por lo que el predictor es más exacto.

3. INDICES DE BIODIVERSIDAD MAS USADOS

3.1 El Índice de Simpson

El índice diversidad de Simpson es una de las medidas más utilizadas. Una búsqueda en Google arroja 28, 600,000 de entradas bajo el descriptor 'Simpson indexdiversity' disminuyendo al hacerla en castellano a 53,000 al usar índice biodiversidad Simpson', de ellas 21,300 corresponden a aplicaciones. Esto da una idea de su popularidad. Veamos los resultados existentes al considerar el problema estadístico. En este estudio el índice es un parámetro que se estima de su estudio como estimador de un parámetro. Este índice toma en consideración el número de especies presentes y su abundancia valorando la proporción de individuos en un y mide su importancia para la diversidad.

Este índice es usado fundamentalmente en cuantificar la biodiversidad de un hábitat por lo que es popular en estudios para fijar el perfil de la biodiversidad. Este fue propuesto por Simpson en 1949. Desde entonces se ha hecho variaciones del mismo.

Sea N_i el número de individuos de la subespecie 'i' en el sitio analizado y N el total de los existentes.

$$\lambda_{Simp} = \sum_{i=1}^k \{1 - \pi_i\} \pi_i = 1 - \sum_{i=1}^k \pi_i^2$$

Se han propuesto varias variantes del índice como el índice de diversidad de Simpson

$$1 - \lambda_{Simp} = \sum_{i=1}^k \pi_i^2$$

Este puede verse como la probabilidad de que dos individuos seleccionados aleatoriamente en un área pertenezcan a dos diferentes subespecies. Este es conocido como índice de concentración de Herfindall que es usado comúnmente en economía en estudios descriptivos. Aquí tiene sentido hablar de dominancia de una especie. Ver Bruckman (1998 y 1999).

El índice recíproco de Simpson es $1/\lambda_{Simp}$ el que valora el número de subespecies igualmente comunes que producirían el valor observado del índice de Simpson. Este índice es influenciado por los parámetros por ciento de cada especie presente y su riqueza. Para una cierta riqueza el índice decrece cuando los porcentos de cada especies tiende a ser igualitaria ($\pi_i=1/K$). El investigador tiene que observar los patrones existentes con cuidado para interpretar el significado de los valores del índice.

Es usual encontrar en la literatura un estimador ingenuo para el índice de Simpson.

$$\hat{\lambda}_{Simp} = 1 - \sum_{i=1}^k \left(\frac{n_i}{n}\right)^2$$

el sesgo de este estimador es $bias(\hat{\lambda}_{Simp}) = -\sum_{i=1}^k \frac{\pi_i(1-\pi_i)}{n}$

Note que el sesgo equivale a la suma de las $var\{n_i\}$, Bouza y Schubert (2003) hicieron la corrección del estimador y calcularon su varianza. Gimaret-Carpentier et al (1998) en un estudio de diversidad además de proponer un estimador insesgado y una varianza basada en métodos no paramétricos. Bouza-Covarrubias (200) analizaron el

comportamiento de $\hat{\lambda}_{Simp(MAScr)} = 1 - \sum_{i=1}^k \frac{n_i(n_i - 1)}{n(n-1)}$

$$Var\left\{\hat{\lambda}_{Simp(MAScr)}\right\} = \frac{2}{n(n-1)} \left[\sum_{i=1}^k \pi_i^2 + 2(n-2)\pi_i^3 - (2n-3)\left(\sum_{i=1}^k \pi_i^2\right)^2 \right]$$

Un estimador de esta es

$$\begin{aligned} & \hat{Var}_1\left\{\hat{\lambda}_{Simp(MAScr)}\right\} = \\ & = \frac{2}{n(n-1)^2(n-2)(n-3)} \left[\sum_{i=1}^k \left\{ 2n^2(n^2 + n - 3)\hat{\pi}_i^3 - n(5n^2 - 3n - 3)\hat{\pi}_i^2 + 3n(n-1)\hat{\pi}_i \right\} - \right. \\ & \quad \left. - n^3(2n-3)\left(\sum_{i=1}^k \hat{\pi}_i^2\right)^2 + 2(2n-3)\sum_{i<j} n^2 \hat{\pi}_i^2 \hat{\pi}_j + n^2 \hat{\pi}_i \hat{\pi}_j^2 - n\hat{\pi}_i \hat{\pi}_j \right] \\ & = \frac{2}{n} \left[\sum \left\{ C_1 \hat{\pi}_i^3 - C_2 \hat{\pi}_i^2 + C_3 \hat{\pi}_i \right\} - C_4 \left(\sum_{i=1}^k \hat{\pi}_i^2\right)^2 + \sum_{i<j} C_5 \hat{\pi}_i^2 \hat{\pi}_j + C_5 \hat{\pi}_i \hat{\pi}_j^2 - C_6 \hat{\pi}_i \hat{\pi}_j \right] \end{aligned}$$

Podemos un estimador sesgado de la varianza que utiliza una fórmula más sencilla. La condición de insesgadez se pierde pero este es consistente. Por su parte un estimador consistente de la varianza del estimador de Simpson es $\hat{Var}_2\left\{\hat{\lambda}_{Simp}\right\}$

$$= \frac{2}{n(n-1)} \left(\sum_{i=1}^k \hat{\pi}_i^2 + 2(n-1)\hat{\pi}_i^3 - (2n-3)\left(\sum_{i=1}^k \hat{\pi}_i^2\right)^2 \right)$$

en donde el número de especies k es conocida, las $\hat{\pi}_i = \frac{n_i}{n}$ dentro del MASCR.

Estos resultados se extienden al establecer que un estimador asintóticamente insesgado es:

$$\hat{Var}_3(\hat{\lambda}_{Simp}) = \frac{2}{n} \left[\sum_{i=1}^k \left(C_1 \hat{\pi}_i^3 - C_2 \hat{\pi}_i^2 + C_3 \hat{\pi}_i \right) - C_4 \left(\sum_{i=1}^k \hat{\pi}_i^2\right)^2 + \sum_{i<j} C_5 \hat{\pi}_i^2 \hat{\pi}_j + C_5 \hat{\pi}_i \hat{\pi}_j^2 - C_6 \hat{\pi}_i \hat{\pi}_j \right]$$

Bouza-Covarrubias (2005) apuntaron que a medida que crece n los productos cruzados se acercan a cero, lo que podría sugerir que la importancia de la interacción entre las especies no tiene un peso importante en la estimación de la varianza si el número de observaciones es grande.

Es usual abordar la estimación de los índices de diversidad para el caso de m sitios de muestreo donde $m > 1$ de un total de M en la población ya que como se apuntó antes los investigadores en ecología realizan muestreos en sus trabajos. Las estimaciones condicionadas a un tamaño fijo es decir las n_j serán consideradas como fijas ($j=1,2,\dots,m$) por lo que la suma de los individuos en la muestra n^* será fijo. La distribución de las especies en cada sitio de muestreo, como usamos muestreo aleatorio simple con reemplazo, es válido usar la distribución Binomial de estos con parámetros n_j y π_i

El estimador ($\hat{\lambda}_{SimpSep(MAScr)}$) y el conjugado ($\hat{\lambda}_{SimpConj(MAScr)}$) fueron desarrollados tomando

$$n_{.j} = \sum_{i=1}^m n_{ij} \text{ y } n^* = \sum_{j=1}^m n_{.j}$$

Derivándose

$$\hat{\lambda}_{SimpSep(MAScr)} = \sum_{j=1}^m \frac{\hat{\lambda}_j}{m}$$

Cuyo error es

$$Var\left\{\hat{\lambda}_{SimpSep(MAScr)}\right\} = \frac{1}{m^2} \sum_{j=1}^m Var\left\{\hat{\lambda}_j\right\}$$

donde los $\hat{\lambda}_j$ son los estimadores para un sitio j .

El estimador Conjugado es el más utilizado por los investigadores en ecología cuando desean caracterizar una población ecológica, siguiendo la idea de conformar una muestra grande con pequeñas muestra, lo que puede representar un problema si no se toma en cuenta los supuestos del MAS. Para estimar de manera conjugada el índice de

Simpson se desarrolló el estimador $\hat{\lambda}_{SimpConj(MAScr)} = 1 - \sum_{i=1}^k \frac{n_i^*(n_i^* - 1)}{n^*(n^* - 1)}$

cuyo error es

$$Var\left\{\hat{\lambda}_{SimpConj(MAScr)}\right\} = \frac{2}{n^*(n^* - 1)} \sum_{i=1}^k \pi_i^2 + 2(n^* - 2)\pi_i^3 - (2n^* - 3)\left(\sum_{i=1}^k \pi_i^2\right)^2$$

El estimador de este error es

$$\begin{aligned} & Var\left\{\hat{\lambda}_{SimpConj(MAScr)}\right\} = \\ & = \frac{2}{n^*(n^* - 1)^2(n^* - 2)(n^* - 3)} \left[\sum_{i=1}^k \left\{ 2n^{*2}(n^{*2} + n^* - 3)\hat{\pi}_i^3 - n^*(5n^{*2} - 3n^* - 3)\hat{\pi}_i^2 + 3n^*(n^* - 1)\hat{\pi}_i \right\} \right. \\ & \left. - n^{*3}(2n^* - 3)\left(\sum_{i=1}^k \hat{\pi}_i^2\right)^2 + 2(2n^* - 3)\sum_{i < j} n^{*2} \hat{\pi}_i^2 \hat{\pi}_j + n^{*2} \hat{\pi}_i \hat{\pi}_j - n^* \hat{\pi}_i \hat{\pi}_j \right] \end{aligned}$$

Tomando el vector paramétrico $\pi(U)=(\pi_1, \dots, \pi_K)^T$ como aleatorio al considerar el MSP

$$\wp: E_{\wp}(\pi_i)=\theta_i, \quad V_{\wp}(\pi_i)=\sigma_i^2, \quad \forall i, i'=1, \dots, K;$$

$$\text{Cov}_{\wp}((\pi_i, \pi_{i'})=\rho\sigma_i\sigma_{i'}=\sigma_{ii'} \quad \forall i \neq i', i, i'=1, \dots, K.$$

Note que estos supuestos nos permite utilizar los resultados asociados a los momentos de una Multinomial. Después de hacer las substitutiones correspondientes y agrupado convenientemente y definiendo los coeficientes

$$\psi(j)=\prod_{i=0}^{j-1}(N-j), \quad j=1, \dots, 4$$

El error esperado bajo este MSP es

$$\begin{aligned} E_{\wp}V\{L_S\} &= \frac{2}{n(n-1)} \left[\sum_{i=1}^k \sigma_i^2 + \left(\frac{2n(4N-3)}{N^4} \right) \Psi(1) \sum_{i=1}^k \theta_i \right] + \\ &+ \frac{2}{n(n-1)} \left[\left(1 + \frac{4n-N-2}{N^4} \Psi(2) \right) \sum_{i=1}^k \theta_i^2 \right] + \\ &+ \frac{2}{n(n-1)} \left[\frac{2n(N-1)-4N+3}{N^4} \Psi(3) \sum_{i=1}^k \theta_i^3 - \frac{(2n-3)}{N^4} \Psi(4) \sum_{i=1}^k \theta_i^4 \right] - \\ &- \frac{2}{n(n-1)} \left[\frac{(2n-3)}{N^4} \left(\Psi(4) \sum_{i \neq i'} \theta_i^2 \theta_{i'}^2 + \Psi(3) \sum_{i \neq i'} \theta_i \theta_{i'} (\theta_i + \theta_{i'}) + \Psi(2) \sum_{i \neq i'} \theta_i \theta_{i'} \right) \right] \end{aligned}$$

A partir de este resultado son deducidos con facilidad los errores bajo la selección de una muestra simple aleatoria con reemplazo de m sitios de muestreo. Tenemos dos predictores alternativos.

Para el índice separado de Simpson el error de este bajo \wp es

$$\begin{aligned} E_{\wp}V\left\{\hat{\lambda}_{SimpSep(MAScr)}\right\} &= \frac{1}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \left[\sum_{i=1}^k \sigma_i^2 + \left(\frac{2n(j)(4N-3)}{N^4} \right) \Psi(1) \sum_{i=1}^k \theta_i \right] \right\} + \\ &+ \frac{1}{m^2} \sum_{j=1}^m \frac{2}{n(j)(n(j)-1)} \left(1 + \frac{4n(j)-N-2}{N^4} \Psi(2) \right) \sum_{i=1}^k \theta_i^2 + \\ &+ \frac{1}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \left[\frac{2n(j)(N-1)-4N+3}{N^4} \Psi(3) \sum_{i=1}^k \theta_i^3 - \frac{(2n(j)-3)}{N^4} \Psi(4) \sum_{i=1}^k \theta_i^4 \right] \right\} \\ &- \frac{1}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \frac{(2n(j)-3)}{N^4} \left(\Psi(4) \sum_{i \neq i'} \theta_i^2 \theta_{i'}^2 + \Psi(3) \sum_{i \neq i'} \theta_i \theta_{i'} (\theta_i + \theta_{i'}) \right) \right\} \\ &- \frac{1}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \frac{(2n(j)-3)}{N^4} \Psi(2) \sum_{i \neq i'} \theta_i \theta_{i'} \right\} \end{aligned}$$

El resultado para el estimador combinado

$$V\left\{\hat{\lambda}_{SimpConj(MAScr)}\right\} = \frac{2}{n^*(n^*-1)} \sum_{i=1}^k \pi_i^2 + 2(n^*-2)\pi_i^3 - (2n^*-3)\left(\sum_{i=1}^k \pi_i^2\right)^2$$

y su esperanza bajo el modelo es

$$\begin{aligned} E_{\varphi} V\left\{\hat{\lambda}_{SimpConj(MAScr)}\right\} &= \frac{2}{n^*(n^*-1)} \left[\sum_{i=1}^k \sigma_i^2 + \left(\frac{2n^*(4N-3)}{N^4}\right) \Psi(1) \sum_{i=1}^k \theta_i + \right] + \\ &+ \frac{2}{n^*(n^*-1)} \left[\left(1 + \frac{4n^*-N-2}{N^4} \Psi(2)\right) \sum_{i=1}^k \theta_i^2 + \frac{2n^*(N-1)-4N+3}{N^4} \Psi(3) \sum_{i=1}^k \theta_i^3 \right] - \\ &- \frac{2}{n^*(n^*-1)} \left[\frac{(2n^*-3)}{N^4} \left(\Psi(4) \left(\sum_{i \neq i'} \theta_i^2 \theta_{i'}^2 - \sum_{i=1}^k \theta_i^4 \right) + \Psi(3) \sum_{i \neq i'} \theta_i \theta_{i'} (\theta_i + \theta_{i'}) \right) \right] - \\ &- \frac{2}{n^*(n^*-1)} \left(\frac{(2n^*-3)}{N^4} \right) \Psi(2) \sum_{i \neq i'} \theta_i \theta_{i'} \end{aligned}$$

Consideremos el MSP caracterizado por las distribuciones utilizadas en la sección anterior

$$\wp(B): E_{\varphi}(\pi_i) = \theta_i, \quad V_{\varphi}(\pi_i) = \theta_i(1-\theta_i)/N, \quad \forall i, = 1, \dots, K.$$

Los errores esperados se expresan como

$$\begin{aligned} E_{\varphi} V\left\{\hat{\lambda}_{SimpConj(MAScr)_{SS}} | B\right\} &= \frac{1}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \left[\sum_{i=1}^k \frac{\theta_i + (N-1)\theta_i^2}{N} \right] \right\} + \\ &+ \frac{1}{m^2} \sum_{j=1}^m \frac{2}{n(j)(n(j)-1)} \left[\left(\frac{4n(j)-N-2}{N^4} \Psi(2) \right) \sum_{i=1}^k \theta_i^2 + \left(\frac{2n(j)(4N-3)}{N^4} \right) \Psi(1) \sum_{i=1}^k \theta_i \right] \\ &+ \frac{1}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \left[\frac{2n(j)(N-1)-4N+3}{N^4} \Psi(3) \sum_{i=1}^k \theta_i^3 \right] \right\} - \\ &- \frac{1}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \left(\frac{(2n(j)-3)}{N^4} \right) \left(\Psi(2) \sum_{i \neq i'} \theta_i \theta_{i'} - \Psi(4) \sum_{i=1}^k \theta_i^4 \right) \right\} - \\ &- \frac{1}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \left(\frac{(2n(j)-3)}{N^4} \right) \left(\Psi(4) \sum_{i \neq i'} \theta_i^2 \theta_{i'}^2 + \Psi(3) \sum_{i \neq i'} \theta_i \theta_{i'} (\theta_i + \theta_{i'}) \right) \right\} \end{aligned}$$

y

$$\begin{aligned}
E_{\varphi} V \left\{ \hat{\lambda}_{SimpConj(MAScr)} | B \right\} &= \frac{2}{n^*(n^*-1)} \left[\sum_{i=1}^k \frac{\theta_i + (N-1)\theta_i^2}{N} \sigma_i^2 + \left(\frac{2n^*(4N-3)}{N^4} \right) \Psi(1) \sum_{i=1}^k \theta_i \right] + \\
&+ \frac{2}{n^*(n^*-1)} \left(\frac{4n^*-N-2}{N^4} \Psi(2) \right) \sum_{i=1}^k \theta_i^2 + \\
&+ \frac{2}{n^*(n^*-1)} \left[\frac{2n^*(N-1)-4N+3}{N^4} \Psi(3) \sum_{i=1}^k \theta_i^3 - \frac{(2n^*-3)}{N^4} \Psi(4) \left(\sum_{i=1}^k \theta_i^4 - \sum_{i \neq i'} \theta_i^2 \theta_{i'}^2 \right) \right] \\
&- \frac{2}{n^*(n^*-1)} \left[\frac{(2n^*-3)}{N^4} \left(\Psi(3) \sum_{i \neq i'} \theta_i \theta_{i'} (\theta_i + \theta_{i'}) + \Psi(2) \sum_{i \neq i'} \theta_i \theta_{i'} \right) \right]
\end{aligned}$$

Cuando N es suficientemente grande para aceptar que $Hn(j)/N \cong 0$ para $H < 5$ estos errores son aproximados adecuadamente por

$$E_{\varphi} V \{ L_{SS} | B \} \cong \frac{1}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \left[(2n(j)-4) \sum_{i=1}^k \theta_i^3 - (2n(j)-3) \left(\sum_{i=1}^k \theta_i^2 \right)^2 \right] \right\}$$

cuando usamos el estimador separado y por

$$E_{\varphi} V \{ L_{SC} | B \} \cong \left\{ \frac{2}{n^*(n^*-1)} \left[(2n(j)-4) \sum_{i=1}^k \theta_i^3 - (2n(j)-3) \left(\sum_{i=1}^k \theta_i^2 \right)^2 \right] \right\}$$

Para el caso del MSP

$$\wp(U): E_{\varphi}(\pi_i) = 1/2, \quad V_{\varphi}(\pi_i) = 1/12, \quad \forall i, = 1, \dots, K.$$

se obtiene

$$\begin{aligned}
E_{\varphi} V \{ L_{SS} | U \} &= \frac{k}{m^2} \sum_{j=1}^m \left\{ \frac{2}{n(j)(n(j)-1)} \left(\frac{1}{2N} + \frac{(N-1)}{4N} \right) \left(\frac{2n(j)(4N-3)}{2N^4} \right) \Psi(1)_i \right\} + \\
&+ \frac{1}{m^2} \sum_{j=1}^m \left(\frac{2}{n(j)(n(j)-1)} \right) \left(\frac{2n(j)(N-1)-4N+3}{8N^4} \right) k \Psi(3) - \frac{(2n(j)-3)}{16N^4} k \Psi(4) - \\
&+ \frac{k}{m^2} \sum_{j=1}^m \left(\frac{4n(j)-N-2}{4N^4} \Psi(2) \right) - \left(\frac{2}{n(j)(n(j)-1)} \right) \left[\frac{(2n(j)-3)}{N^4} \left(\frac{\Psi(4)}{16} + \frac{\Psi(3)}{16} \right) \right] \cong \\
&\cong \frac{k}{m^2} \sum_{j=1}^m \left(\frac{n(j)-1}{2n(j)(n(j)-1)} \right) + \left(\frac{2n(j)-3}{8n(j)(n(j)-1)} \right)
\end{aligned}$$

y

$$\begin{aligned}
E_{\wp} V\{L_{SC}|U\} &= \frac{2\Psi(1)}{n^*(n^*-1)} \left(\frac{1}{2N} + \frac{N-1}{4N} \right) \frac{2kn^*(4N-3)}{2N^4} + \frac{k\Psi(2)}{4N^4} (4n^*-N-2) + \\
&+ \frac{2k}{n^*(n^*-1)} \left[\left(\frac{2n^*(N-1)-4N+3}{8N^4} \right) \Psi(3) - \left(\frac{(2n^*-3)k\Psi(4)}{16N^4} \right) \right] - \\
&- \frac{2}{n^*(n^*-1)} \left[\left(\frac{(2n^*-3)}{N^4} \right) \frac{\Psi(3)+\Psi(4)}{16} \right] \cong k \frac{n^*-2}{2n^*(n^*-1)} - \frac{2n^*-3}{8n^*(n^*-1)}
\end{aligned}$$

Mientras que para

$$\wp(\text{Be}): E_{\wp}(\pi_i) = \alpha_i(\alpha_i+1)\beta_i, \quad V_{\wp}(\pi_i) = \alpha_i\beta_i(\alpha_i+\beta_i)^{-2}(\alpha_i+\beta_i+1) \quad \forall i, =1, \dots, K$$

estos errores esperados son aproximados por

$$\begin{aligned}
E_{\wp} V\left\{ \hat{\lambda}_{\text{SimpConj(MAScr)}} | \text{Be} \right\} &\cong \sum_{j=1}^m \frac{2}{n(j)(n(j)-1)} \left(\frac{\alpha_i\beta_i}{(\alpha_i+\beta_i)(\alpha_i+\beta_i+1)} + \left(\frac{\alpha_i}{(\alpha_i+\beta_i)} \right)^2 \right) + \\
&+ (2n(j)-4) \sum_{j=1}^m \left(\frac{\alpha_i}{(\alpha_i+\beta_i)} \right)_i^3 - (2n(j)-3) \left(\sum_{i=1}^k \left[\frac{\alpha_i}{(\alpha_i+\beta_i)} \right]^2 \right)_i^2
\end{aligned}$$

y

$$\begin{aligned}
E_{\wp} V\left\{ \hat{\lambda}_{\text{SimpConj(MAScr)}} | \text{Be} \right\} &\cong \frac{2}{n^*(n^*-1)} \left(\frac{\alpha_i\beta_i}{(\alpha_i+\beta_i)(\alpha_i+\beta_i+1)} + \left(\frac{\alpha_i}{(\alpha_i+\beta_i)} \right)^2 \right) + \\
&(2n^*-4) \left(\frac{\alpha_i}{(\alpha_i+\beta_i)} \right)_i^3 - (2n^*-3) \left(\sum_{i=1}^k \left[\frac{\alpha_i}{(\alpha_i+\beta_i)} \right]^2 \right)_i^2
\end{aligned}$$

3.2 El Índice de Shannon

El índice de Shannon es a veces citado como de Shannon-Wiener (o de Shannon-Weaver) y es uno de los más populares junto con el de Simpson. Usando en Google aparecen 157,000 citas bajo el descriptor 'indexbiodiversity Shannon', 15,700 para 'indexbiodiversity Shannon-Wiener' y 8,920 para 'indexbiodiversityshannon-weaver'. Claude Shannon publicó un par de trabajos seminales para la Teoría de Información. Por su parte Weaver en su rol de administrador de la Rockefeller Foundation fue un mecenas de esta nueva teoría dándole soporte a las investigaciones. Por su parte Wiener colaboró a la divulgación de esta teoría a través de su trabajo en cibernética. Ambos son considerados como personajes importantes al considerar el desarrollo de la Teoría de la Información pues su labor se jugó un rol decisivo después de la Segunda Guerra Mundial. Esta teoría es una rama de las, matemáticas aplicadas y es muy usada en ingeniería eléctrica. Una vez fijados los conceptos fundamentales esta teoría es usada en enfoques diversos. Uno de sus usos está en la estadística inferencial, el procesamiento del lenguaje natural, en criptografía, neurobiología análisis de datos y otras ramas de carácter aplicado. Su concepto clave es la entropía.

La matemática que soporta esta teoría fue desarrollada para la termodinámica. Por tanto se vincula a labor de J. Willard Gibbs y Ludwig Eduard Boltzmann fundadores de la estadística mecánica y termodinámica. Estas conexiones fueron estudiadas por Rolf William Landauer el que demostró que cuando la información se pierde en un circuito irreversible la información se convierte en entropía y la energía asociada se disipa como calor.

En el caso de la biodiversidad la entropía de la información es asociada a la distribución. Las especies se consideran símbolos y su frecuencia relativa como probabilidades. La bondad dada a este índice es que toma en cuenta el número de especies y la rareza.

Tomando:

N_i = número de individuos de la especie i en la población (abundancia de la especie i).

K = número de especies (riqueza)

$N = \sum_{i=1}^K N_i$ = Total de individuos en la población

$\pi_i = N_i / N$ = abundancia relativa de la especie (probabilidad de observar un individuo de la especie al hacer una selección aleatoria)

El índice de biodiversidad de Shannon es $S = -\sum_{i=1}^K \pi_i \ln(\pi_i)$. En muchas ocasiones se utiliza el índice $S' = S + [(K-1)/2N]$. Note que para cualquier número de especies fijado su máximo ocurre cuando el número en cada especie es igual en el que se tiene que $S_{\max} = \ln(K)$.

La estructura de este índice es no lineal en los p_i 's, lo que dificulta su estudio mediante las técnicas estadísticas clásicas. Sin embargo dada la base teórica, proveniente de la Teoría de la Información, que le soporta es muy usado.

Al tomar una muestra de tamaño n estimaríamos el índice de Shannon, bajo el supuesto de que K es conocido, en un sitio de muestreo al evaluar el número de individuos de cada especie i en la muestra, n_i , entre los n observados

$$\hat{S} = - \sum_{i=1}^K p_i \ln(p_i)$$

donde $p_i = n_i/n$. El vector $\underline{n} = (n_1, \dots, n_K)$ se distribuye Multinomial con parámetros n y $\underline{\pi} = (\pi_1, \dots, \pi_K)$ por lo que este estimador es Máximo Verosímil por serlo el vector $\underline{p} = (p_1, \dots, p_K)$ para $\underline{\pi}$. Este tiene un sesgo negativo y su esperanza es $E(\hat{S}) = S - [(K-1)/2N] + O(n^{-2})$.

Estos resultados fueron deducidos por Pielou (1975). Note que es necesario que $n > K$ y que usando

$$\hat{S}' = \hat{S} + [(K-1)/2N] + O(n^{-2}).$$

tenemos un estimador asintóticamente insesgado.

Si una especie no es observada $n_i=0$ y el modelo no es válido. Un estimador fue propuesto por Chao et al. (2002) basado en la estimación de Horvitz-Thomson. Este es

$$\hat{S}(HT) = - \sum_{i=1}^K \frac{p_i \ln(p_i)}{1 - (1 - p_i)^n} I(A(i))$$

donde

$$I(A(i)) = \begin{cases} 1 & \text{si la especie } i \text{ está en la muestra} \\ 0 & \text{sino} \end{cases}$$

Estos propusieron usar el cubrimiento de la muestra

$$C = \sum_{i=1}^K \pi_i I(A(i))$$

Este mide la fracción del total de la abundancia de las especies en la muestra, por lo que es la probabilidad condicional de descubrir una nueva especie, ver Good (1953). Entonces el modelo Multinomial tiene parámetros n y

$$\pi_i^* = \frac{\pi_i}{C}, i = 1, \dots, K$$

El estimador que proponen es

$$\hat{S}(HT) = - \sum_{i=1}^K \frac{\hat{C} p_i \ln(\hat{C} p_i)}{1 - (1 - \hat{C} p_i)^2} I(A(i))$$

Evidentemente en ningún caso los estimadores son insesgados. Se clama por la consistencia o la insesgader asintótica.

En los modelos que le usan se hace lo que podríamos considerar una especie de estimador del tipo conjugado, pues al tomar la muestra de m sitios, que lo usual los conteos sean

$$n_i = \sum_{j=1}^m n_{ij}, \text{ donde}$$

n_{ij} es el número de ejemplares de la especie i observados en la muestra j

Si decidimos usar el método separado al tomar m sitios de muestreo se computaría la media de los estimadores, digamos

$$\bar{S}(s) = \frac{1}{m} \sum_{j=1}^m \hat{S}(j), \quad \hat{S}(j) = \text{estimador de } S \text{ en la muestra } j\text{'ésima}$$

$$p_i(j) = n_i(j)/n_j$$

En este caso $n_i(j)$ es el número de individuos de la especie i en la muestra j y n_j el tamaño de esta.

Contemporáneamente a través de métodos de remuestreo como Bootstrap y Jackknife se hacen los estudios basados en este índice.

3.3 El Índice de Fager

El índice inicialmente propuesto por Fager (1972) es

$$:\lambda''_F = [N(K+1) - J(K-J)]/2 - \sum_{i=1}^K N_i R_i = \lambda^*_0 - \sum_{i=1}^K N_i R_i$$

Donde $J \in [0, K)$ es un entero y R_1, \dots, R_K son rangos asignados a las especies de interés en orden decreciente a su importancia. Esto es fijado de antemano por el ecólogo. Este llamó la atención por sus propiedades desde sus inicios. Referencias al mismo aparecen en la literatura con frecuencia en las aplicaciones, ver por ejemplo Oehl et al (2003).

En el orden más teórico no son abundantes las contribuciones. Esto ocurrió en los años cercanos a la propuesta del índice. Podemos ver que Hayes (1978) desarrolló un estudio de Monte Carlo para reproducir muestras de un grupo original de mediciones. Lo que llamó índice de Fager del grupo recurrente se calculó y se analizó el efecto del número de muestras. La existencia de un sesgo negativo fue reportado cuando el tamaño de la muestra era pequeño, sin hacer una referencia a propiedades teóricas del índice como estadístico. Bouza-Schubert [2004] propusieron la simple transformación del índice:

$$\lambda_F = [N(K+1) - J(K-J)]/2N - \sum_{i=1}^K W_i R_i = \lambda_0 - \sum_{i=1}^K W_i R_i$$

donde

$$\lambda_0 = [N(K+1) - J(K-J)]/2N$$

Esto permite obviar el desconocimiento de N pues en particular esto nos permite considerar que

$$\lim_{N \rightarrow \infty} [N(K+1) - J(K-J)]/2N = [K+1]/2.$$

En este modelo tomamos un sitio aleatoriamente y observamos una estructura muestral (n_1, \dots, n_k) , donde n_j es el número de individuo de U_j en el sitio muestreado. Note que la distribución de (n_1, \dots, n_k) es una Multinomial y que el número de individuos en la muestra es $n = \sum_{i=1}^K n_i$. Por tanto $E(n_i) = nW_i$ y $p_i = n_i/n$ estima la proporción W_i de la especie i . Fager (1972) asumió que $R_i = \text{Rank}(N_i)$ es fijado por el ecólogo de antemano. Nosotros asumiremos, de ser necesario, además que $R_i \cong \text{Rank}(nW_i)$ en forma tal que la relación $R_i \cong \text{Rank}(n_i) = r_i$ sea aceptable. Entonces podemos escribir

$$\lambda^\circ_F = \lambda_0 - \sum_{i=1}^K p_i r_i$$

Entonces si usamos rangos basados en el rango asignado en función de información muestral $E(\lambda^\circ_F) \cong \lambda_F$ y su varianza es:

$$V(\lambda^{\circ}_F) = \sum_{i=1}^K R_i^2 W_i (1-W_i) / n = V^2 V_2 / n$$

Cuyo estimador insesgado es, ver Bouza-Schubert (2003):

$$V^{\circ}(\lambda^{\circ}_F) = \sum_{i=1}^K R_i^2 p_i (1-p_i) / n.$$

Muchas de las investigaciones en ecología tienden a caracterizar una comunidad y en la mayoría de las ocasiones se desea obtener información de áreas o superficies muy grandes. Esto determina la necesidad de utilizar muestras. Kepton (2002) señala la necesidad de dar una interpretación teórica a los índices como medidas de diversidad a partir de estudios empíricos. El problema de estimar índices se enfrenta a la necesidad de desarrollar una teoría de muestreo particular pues no se cuenta con un marco muestra.

En el marco usualmente utilizado en la teoría de encuestas se utiliza una sucesión de muestras aleatoriamente independientes $s_h, h=1, \dots, m$, de tamaño no necesariamente igual. Entonces tendremos una sucesión de variables Multinomiales $n_h = (n_{1h}, \dots, n_{Kh})$, $h=1, \dots, m$ y

$$\lambda_{F,h} = \sum_{i=1}^K p_{i,h} R_i = \sum_{i=1}^K R_i \sum_{h=1}^m n_{ih} / n_h, \quad h=1, \dots, m$$

y estima insesgradamente el índice de Fager utilizando a

$$\lambda^*_{F} = \lambda_0 - \sum_{i=1}^K \lambda_{F,h} / m.$$

Ya que $n_{ih}/n_h = p_{ih}$ estima insesgradamente a W_i . Estos resultados permiten dar como válida la siguiente Entonces al tomar m muestras independientes mediante muestreo simple aleatorio

$$\lambda^*_{F} = \lambda_0 - \sum_{i=1}^K \lambda_{F,h} / m.$$

es un estimador insesgado de λ_F donde λ_0 es fijado por [2.1] y $\lambda_{F,h}$ por [2.4]. Su varianza es

$$V[\lambda^*_{F}] = V^2 \sum_{h=1}^m n^{-1}_{,h} / m$$

Un estimador insesgado de ella es

$$V^{\circ}(\lambda^*_{F}) = \sum_{i=1}^K R_i^2 \sum_{h=1}^m p_{ih} (1-p_{ih}) / n_h m$$

Si analizamos la estructura del índice de Fager es claro que esta muestra similitudes con los estadísticos lineales de rango. Sin embargo K no es el tamaño de una muestra sino un parámetro. Esto hace que los teoremas que aparecen en la literatura para establecer la convergencia de su distribución no sean utilizables directamente. Un expresión condensada de los teorías que pueden ser vinculados con tales estadísticos es el que brindamos a continuación y que se basa en Chernoff et al. [1967] usando la notación de Jurečková-Sen [1996] y Friedst-Gray [1997].

Teorema 3.1. [Convergencia de la distribución de estadísticos lineales de rango].

Sean

- X_1, \dots, X_m una sucesión de variables aleatorias de una función de distribución F .
- $X_{(1)}, \dots, X_{(m)}$ la correspondiente de los estadísticos de orden.
- Una función continua $J:]0, 1[\rightarrow \mathfrak{R}$ tal que $|J(u)|$ y $|F^{-1}(u)|$ están acotadas por $au^{2-r/2r}(1-u)^{1-2/2s}$ para $a, r, s > 0$ y $0 < u < 1$. Entonces $T_m = \sum_{i=1}^m J[u/(m+1)]X_{(i)}$ se distribuye asintóticamente $N\{\mu_m, \sigma^2(J, F)\}$ donde
 - $\mu_m = \int_0^1 J(u) F^{-1}(u) du$
 - $\sigma^2(J, F) = \int_0^1 [J(u)F^{-1}(u)]^2 du - \left[\int_0^1 J(u)F^{-1}(u) du \right]^2$

Ideas sobre su demostración puede ser obtenidas en Ferguson [1999] y Caaperà-Cutsem [1988] por ejemplo.

La estructura de la componente aleatoria de [2.5] es similar a los de los estadísticos T_m llamados de " tipo L-estimador" pues podemos escribirle como $\sum_{i=1}^K \lambda_{F,h}/m = \sum_{h=1}^m [1/n_{i,h}] [\sum_{i=1}^K n_{i,h} R_i] / m$, haciendo $J[h/(m+1)] = 1/n_{i,h}$. Para satisfacer las hipótesis 1 y 2 del teorema necesitaríamos que $\sum_{i=1}^K n_{i,h} R_i = X_{(h)}$

lo que lograríamos al hacer una ordenación de los subíndices basados en la muestra.

Esto es factible pero la función de pesos no es constante pro que la hipótesis 3 no se cumple. Sin embargo esto nos hace suponer que el estimador puede seguir una distribución asintótica normal.

Siguiendo otro camino como n_{ih} sigue una distribución la Binomial $B\{n_h, W_h\}$ por lo que si n_h es suficientemente grande p_{ih} sigue asintóticamente una distribución normal $N\{W_i, W_i[1-W_i]/n_h\}$. Esto es generalmente valido pues al trabajar sobre U las probabilidades W_i son constantes para cada especie y las n_h 's son suficientemente grandes para aceptar la aproximación normal de la Binomial paara cada i y h . Como los rangos son fijados usando información a priori sobre la variable $Y_{ih} = R_i p_{ih}$ se distribuye asintóticamente $N\{R_i W_i, R_i^2 W_h [1-W_i] / n_h\}$.

Sea $Y_{ih}^* = Y_{ih} - R_i W_i$ y $\sigma_i^2 = R_i^2 W_h [1-W_i]$ y la variable $Z_h = \sum_{i=1}^m Y_{ih}^*$ sigue una distribución

$N\{0, \sum_{i=1}^m \sigma_i^2 / n_h\}$. Al tomar una muestra aleatoria independiente mediante msa

$$Z_m = \sum_{h=1}^m r_{hm} Z_h = \sum_{h=1}^m r_{hm} [\sum_{i=1}^K Y_{ih} - E\{Y_{ih}\}]$$

tiene esperanza igual a cero y como varianza a $V[Z_m] = V^2 [\sum_{h=1}^m 1/r_{hm}] / m^2 = V^2 / m^*$ al tomar $r_{hm} = 1/m$ para todo $h=1, \dots, m$. Como los tamaños de las muestras n_h 's en general son grandes la sumatoria en ese término debe ser pequeña por lo que si $m \rightarrow \infty$ los conjuntos $c[h] = \{Z_h | Z_h \notin [-\epsilon V \sqrt{m}, \epsilon V \sqrt{m}], \forall \epsilon > 0\}$ tienen medida de Lebesgue nula.

Entonces, $\lim_{m \rightarrow \infty} [V/\sqrt{m^*}]^{-1} \sum_{h=1}^m \{E(Z_h^2)I(|Z_h| < \epsilon V\sqrt{m})\} = 0$. Esto no permite argüir, ver Ferguson [1999], que Z_m cumple con las condiciones del Teorema de Lindberg-Feller lo que garantiza la convergencia de $Z_m\sqrt{m^*}/V$ a una distribución normal standard. Estos razonamientos nos llevan al aceptar como válida la siguiente proposición.

Proposición 3.2. La distribución asintótica de $\lambda^*_{F} = \lambda_0 - \sum_{i=1}^K \lambda_{F,h}/m$ es $N\{\lambda_F = \lambda_0 - \sum_{i=1}^K R_i, W_i, V\{Z_m\}$, si $m \rightarrow \infty$.

Condicionando a que el número de individuos en la muestra s es fijo se hace el análisis usual pero es claro que al seleccionar un sitio el número de elementos en él no es conocido, por lo que es una variable aleatoria. Al analizar el problema que nos plantea este hecho se debe buscar un modelo apropiado considerando la existencia de una fuente de variación debida a la aleatoriedad de n . Este fenómeno se modela al definir un cierto modelo probabilístico como generador del número de individuos en la clase de interés. Un acercamiento al problema es considerar que el sitio de muestreo es generado por un modelo en el que se asignan v unidades a la población Q y que este es un valor conocido o conocible. Por ejemplo se puede saber que hay vv unidades de biomasa usando un espectrómetro de masa al analizar fotos de un satélite que dio un mapa de la zona y usar la variable

$$Z_C(u) = 1 \text{ si } u \in C$$

$$Z_C(u) = 0 \text{ en otro caso}$$

Esta es una variable Bernoulli con parámetro P . Entonces tomado C como la población de interés U

$$.n[v] = \sum_{h=1}^v Z_C(u) \text{ se distribuye binomialmente con parámetro } v \text{ y } P.$$

Bouza-Covarrubias tomaron un modelo superpoblacional en él se ligan estos puntos de vista en el análisis. Si consideramos que el espectrómetro evalúa d variables y que $Z_U[u] = g[q_1, \dots, q_d] + \epsilon$

donde ϵ es una variable aleatoria con esperanza cero.

El problema de estimar la biodiversidad lo podemos enfocar en la forma siguiente:

- La naturaleza genera una población Q con v unidades.
- Es de nuestro interés solo las especies en $U \subset Q$.
- Se toma muestras usando un diseño muestral aleatorio y se observa en cada selección $n(v)$ unidades de U con las que se computa un índice de biodiversidad.
- El error de muestreo del índice depende del número de individuos observados en la muestra $n(v)$ que es una variable aleatoria.

-Es de interés estimar el error de estimación pero también conocer su esperanza la que depende del modelo superpoblacional g a través del comportamiento de $n(v)$ que es descrito por un modelo Binomial.

Es claro que dada la naturaleza del problema $n(v)$ sigue una $B[v,P]$ por lo que $E[n(v)|g]=vP$ y $Var=n(v)|g]=vP[1-P]$. Como la población fija es generada por g , al observar una sucesión de realizaciones y dada la convergencia de la Binomial a la normal tenemos, ver Friedst-Gray [1997] que el proceso puede ser descrito asintóticamente por una distribución $N\{vP, vP[1-P]\}$ pues

$$[n(v)-vP]\sqrt{v} \longrightarrow_L N\{0, vP[1-P]\}$$

donde la convergencia es en ley. Para una realización $n(v)$ el error de λ_F es

$$V[\lambda_F]=\sum_{i=1}^K R_i^2 W_i [1-W_i]/n(v).$$

El resultado central del trabajo de referencia es que si v es suficientemente grande para aceptar que

$$[n(v)-vP]\sqrt{v} \longrightarrow_L N\{0, vP[1-P]\} \text{ entonces } E[V(\lambda_F|g)] \cong [vP]^{-1} [1+[1-P]/vP] \sum_{i=1}^K R_i^2 W_i [1-W_i]$$

De lo anterior tenemos que el índice de Fager posee las mejores condiciones desde el punto de vista estadístico para ser usado en las inferencias. Puede notarse que este es menos citado por los especialistas en sus investigaciones aplicadas. Otros se han impuesto por la tradición de su uso, como el de Shannon, pero su significado es confuso a la hora de hacer estudios, como la riqueza, basado en modelos estadísticos. Prefieren hacer análisis sin base probabilística por lo que las inferencias son dudablemente creíbles. Podemos citar algunas aplicaciones donde este es usado. Cuando que los especialistas poseen mayor control sobre el problema el índice Fager es más popular. Por ello podemos verle usado más frecuentemente en algunas aéreas, como la medicina, donde se puede aquilatar mejor la importancia de las especies en el problema bajo estudio.

Esto ha ocurrido por ejemplo: al investigar Albert et al. (2004) la variación de la diversidad clonal en la población de *V. myrtilus* en diferentes Hábitats de una área geográfica y al estudiar el particionamiento de la variabilidad genética del aislamiento de DNA ISOLATION extraído usando el modificaciones del procedimiento CTAB. En estos estudios la concentración de DNA se llevó a cabo mediante fluorometría usando bisbenzimidazol. Ver Labarca- Paigen, 1980). La diversidad genotípica fue estudiada usando 3 índices entre los que se incluye el propuesto por (Fager, 1972)

Bonet et al. (2002) consideraron la biodiversidad de microorganismos existentes en el intestino delgado humano. Hay un reconocimiento reciente de que, tener un cuadro de esta biodiversidad, ayuda a la biología molecular, al acceder directamente a la diversidad bacteriana rDNA, en la comunidad natural del intestino. Esto ha generado estudios directos de DNA PCR sobre la comunidad. Estos son usados crecientemente para dilucidar incógnitas sobre la diversidad de la comunidad dentro de diversos ecosistemas. Tal es el caso de experiencias variadas como las Whitford et al.,(1998)

sobre el rumen, de los digestores anaerobios de Godonet *al.*, (1997), en fuentes termales de Barns *et al.*, (1994), de suelos de Liesack-Stackebrandt, (1992) y en los sedimentos marinos de Devereux-Mundfrom, (1994). En su estudio Bonet et al (2002) examinaron el efecto del número de ciclos de PCR de la estructura inferida de una librería, considerando el comportamiento de métodos matemáticos que permitieron comparar estadísticamente la librería del clonaje de rDNA. Su objetivo era establecer si existían diferencias significativas entre las librerías ampliadas de rDNA, por usar diferentes números de ciclos en el PCR. Utilizaron Bootstrap y pruebas Chi-cuadrado en el análisis de la distribución de clones. En el análisis se incluyó el índice de Fager, considerando que se basa en el número de movidas, para caracterizar los movimientos requeridos para convertir la distribución de clones, observada entre los OTUs, en una distribución adecuada, citando la experiencia de (Fager, 1972). Los intervalos de confianza derivados son más creíblemente aceptables, dada las características del índice como estadístico

Menares- Sepúlveda (2005) estudiaron simultáneamente 80 estaciones oceanográficas distribuidas en 24 transectos perpendiculares a la costa. Se estudió la estructura comunitaria de peces y crustáceos demersales, en la zona centro-sur de Chile. Como base del estudio se considera que las especies marinas presentan afinidad. Estos calcularon los índices de diversidad de Shannon y de Fager entre otros. Usando la idea presentada por Fager (1957) de que valores por encima del 0,5 con respecto al máximo número de afinidades posibles. Utilizaron Análisis de Correlación Canónica utilizando la abundancia relativa normalizada de las especies registradas y que estuvieron presentes a lo menos en un 20% del total de los lances. La similitud faunística se determinó utilizando las especies dominantes determinadas según el método de ordenamiento propuesto por Fager (1957), donde las especies se ordenan (rankean) de crecientemente del 1 al 10 en términos de la abundancia dentro de cada muestra. Estos rangos se suman y se dividen por el número de lances. Según el método Fager, de un total de 96 lances y 38 especies presentes se determinaron las 10 especies más importantes. Esto permitió el ranqueo para el estudio de la biodiversidad marina.

REFERENCIAS

Adams, J. E. and E.D. McCune (1979): Application of the generalised Jackknife to Shannon's measure of information used as an index of diversity. En "Ecological Diversity in Theory and Practice" (J. F. Grassie et al., eds) International Co-operative Pub. House. Fairland, Maryland. 150-165.

Albert, Thierry, Olivier Raspé and Anne-Laure Jacquemart (2004): Clonal diversity and genetic structure in *vaccinium myrtillus* populations from different habitats, Belg. Journ. Bot. 137: 155-162.

Allende, S. C. Bouza, Amita D. Chakrabortya (2013): STUDY OF THE COVERAGE OF PLANTS AND THE EVALUATION OF DEFORESTATION, En DEFORESTATION, CONSERVATION POLICIES, ECONOMIC IMPLICATION AND ENVIRONMENTAL IMPACT. (C. Bouza, EDITOR), 101-114. NOVA PUBLISHERS, N. YORK. ISBN: 978-1-62948-241-5 • e-book ISBN: 978-1-62948-242-2.

Almazán A., Urbán G., González R., Tapia J., Villerías S., Beltrán E., Almazán M. (2003), "Ordenamiento Ecológico Territorial de la Subcuenca del Río San Juan del Estado de Guerrero," **Reporte Técnico de Investigación**, SBEJ-UAGroI

BanjShafiei, A., M. Akbarinia, S. Gholami Jalili, P. Azizi and S. Mohsen Hosseini (2006): Effect of fire on herb layer biodiversity in a temperate forest of Northern Iran. *Pakistan J. of Biological Sciences*. 9.

- Barns, S. M., Fundyga, R. E., Jeffries, M. W. & Pace, N. R. (1994). Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc Natl Acad Sci USA* 91, 1609±1613.
- Bonnet, R., A. Suau, J. I Doré, G. R. Gibson and Matthew D. Collins (2002): Differences in rDNA libraries of faecal bacteria derived from 10- and 25-cycle PCRs. *International Journal of Systematic and Evolutionary Microbiology*, 52, 757–763
- Bouza, C. (1996): Linear rank tests derived from a superpopulation model. *BiometricalJ* , 38, 497-506.
- Bouza, C. (2009): Estudio del cubrimiento vegetal: comparación de alternativas basadas en conjuntos ordenados. Estudio del cubrimiento vegetal: comparación de alternativas. Editorial Universitaria (Cuba): Ciudad de La Habana, ISBN 978-959-16-1083-6
- Bouza C. y D. Covarrubias (2005). ESTIMACIÓN DEL ÍNDICE DE DIVERSIDAD DE SIMPSON EN m SITIOS DE MUESTREO. *Inv. Operacional*, 26, **187-195**
- Bouza C. y D. Covarrubias (2007):.PROPUESTA DE UN MARCO SUPERPOBLACIONAL PARA EL ESTUDIO DE LA DIVERSIDAD MEDIANTE LOS INDICES DE FAGER Y SIMPSON. *REVISTA INVESTIGACIÓN OPERACIONAL*, . 28, 25-37,
- Bouza C. N.y L. Schubert (2003). The Estimation of The Biodiversity and The Characterization Of The Dynamics: An Application To The Study Of A Pest. (2003) *Revista De Matematica E Estatistica*. 21, 85-98/
- Beran, H. (1994): Eine system Konforme Maßzahlbildung für Klassifikatorische Merkmale. *Operations Res. Spektrum*. 16, 81-88.
- Beran, H. (1999): A generalisation and optimisation of a measure of concentration. In “Modelling and Decisions in Economics” (U. Leopold-Wildburger, U. et. al. Eds.) Physika Verlag, Heidelberg.
- BjörnNordén, Heidi Paltto, Frank Götmark, KjellWallin (2007):Indicators of biodiversity, what do they indicate? – Lessons for conservation of cryptogams in oak-rich forest. *Biological Conservation*, 135, 369-379.
- Brained, B. (1982): On the relation between the type-token and species problem. *J. Applied Prob.* 19, 785-793.
- Brown JH, Lomolino MV. (2004) *Biogeography*. 3rd. Sunderland, MA: Sinauer Associates,
- Bruckman, G. (1969). Einige Bemerkungen zur statischen Messung der Konzentration. *Metrika*, 14, 183-213.
- Buckland, S. T., A.E. Magurran, R.E. Green and R.M. Fewster (2005): Monitoring change in biodiversity through composite indices. *Phil. Trans. R. Soc. B* 360, 243–254
- Bunnell F.L., B.G. Dunsworth, D.J. Huggard, and L.L. Kremsater.(2003): . Learning to sustain biological diversity on Weyerhaeuser’s coastal tenure. *The Forest Project*, Weyerhaeuser, Nanaimo, BC.
- Burgmann MA, Lindenmayer DB. (1998) *Conservation Biology for the Australian Environment*. Chipping Norton NSW: Surrey Beatty and Sons Pty, Ltd,
- Caaperáa, Ph. Y B. Van Cutsem (1988) *Asymptotique :Méthodes et Modèles en Statistique Non Paramétrique*. Exposition Fondamental. Bordas, Paris.
- Chao, A. (1981): On estimating the probability of discovering a new species. *Ann. Statistics*, 9, 1339-1342.
- Chao, A. and T. Bunge (2002): Estimating the number of species in a stochastic abundance model. *Biometrics*. 58, 530-539.

- Chao, A. and S. M. Lee (1992): estimating the number of classes in sample coverage. *J. American Stat. Ass.* 87, 210-217.
- Chao, A. and Tsung-Jen Shen (2002): Nonparametric estimation of Shannon's index of biodiversity when there are unseen species in simple. *Environmental and Ecological Statistics*, 10, 429-443.
- Cook, A., Marion, G., Butler, A. & Gibson, G. (2007). Bayesian inference for the spatio-temporal invasion of alien species. *Bulletin of Mathematical Biology* 69(6), 2005-2025.
- Chernoff H., J.L. Gaswirth Y M.V. John. (1967): Distribution Of Linear Combination Of Functions Or Order Statistics With Applications To Estimation. *Ann. Math. Statist.* 38, 52-72.
- Ducluzeau, R. (1988). Role of experimental microbial ecology in gastroenterology. In *Microbial Ecology and Intestinal Infections*, (E. Bergone-Berezin, Editor). , 7-26 Berlin: Springer.
- Efron, B. and R. Tibshirani (1976): Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika*, 63, 435-447.
- Engen, S. (1996): *Stochastic Abundance Models*. Halted Press, N. York.
- Esty, W.W. (1986): The efficiency of Good's nonparametric coverage estimator. *Ann. Of Statistics*. 14, 1-9.
- Fager, E.W. (1957). Determination and analyses of recurrent groups. *Ecology*, 38: 586-595.
- Fager, E. W.(1963): The composition of sea-water comparative and descriptive oceanography. Communities of organism.s. In: M. N. Hill (Ed.), *Thea Sea. Ideas and Observations on Progress in the Study of the Seas. II.* 415-437 .Interscience Publ., Londres
- Fager E.W. (1972): Diversity: a sampling study. *American Naturalist*, 106, 293-310.
- Fager, E.W. & J.A. McGowan. (1963). Zooplankton species groups in the North Pacific. *Science*, 140: 453-460.
- Feller, W. (1966): *An Introduction to Probability Theory and Its Applications*. Vol. 2. J. Wiley, N. York.
- Ferguson T. S. (1996): *A Course in Large Sample Theory*. Chapman and Hall. N. York.
- Friedst B. and Gray L. (1997) *A Modern Approach To Probability Theory*. Birhãuser, Boston.
- Fisher, R.A, A. S. Corbett and S.C.B. Williams (1943): The relation between the number of species and the number of individuals in a random sample of animal population. *J. of Animal E.* 12, 42-58.
- Gimaret-Carpentier C, Pélissier R., Pascal J. and Houllier F. (1998) "Sampling Strategies For The Assessment Of Tree Species Diversity," *Journal Of Vegetation Science*, 9, 161-172.
- Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. (1997). Molecular microbial diversity of an anaerobic digester as determined by small-subunit rDNA sequence analysis. *Appl Environ. Microbi*, 63, 2802-2813.
- Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* 40, 237-264.
- Gotelli, and R.K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4:379-391.
- Johnson N. L. and S. Kotz . (1969), *Distributions in Statistics: Discrete Distributions* John Wiley & Sons, N. York.
- Jurecková, J. y P.K. Sen (19976): *Robust Statistical Procedures: Asymptotic And Interrelations*. J. Wiley.N. York.

- Kempton R. A. (2002) "Species Diversity," Encyclopedia of Environmetrics, 4, 2086-2092. (eds Abbel H. El-Shaarawi and W. Piegorsh) John Wiley & Sons. N. York.
- Kevin J. Gaston and John I. Spicer. 2004. Biodiversity: an introduction, Blackwell Publishing. 2nd Ed.
- Keylock, C. J. (2005): Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. *Oikos*, 109, 203-207.
- Kirchhoff T, Trepl L (2001) Vom Wert der Biodiversität – Über konkurrierende politische Theorien in der Diskussion um Biodiversität. *Zeitschrift für angewandte Umweltforschung Sonderheft 13/2001*: 27-44. S
- Labarca C. and Paigen K., 1980. — A simple, rapid, and sensitive DNA assay procedure. *Anal. Biochem.* 102: 344-352.
- Liesack, W. and Stackebrandt, E. (1992). Occurrence of novel groups of the domain *Bacteria* as revealed by analysis of genetic material isolated from an Australian terrestrial environment. *J Bacteriol* 174, 5072-5078.
- Lindenmayer, D.B. and J.F. Franklin. (2002):. *Conserving forest biodiversity: A comprehensive multiscale approach.* Island Press, Washington, DC.
- Lindenmayer, D.B., C.R. Margules and D.B. Botkin (2000): Indicators of biodiversity for ecologically sustainable forest management. *Conservation Biology* 14(4): 941-950.
- MacKay, David J. C. (2003): Information Theory, Inference, and Learning Algorithms Cambridge: Cambridge University Press.
- Margalef, D.R. 1958. Information theory in ecology. *Gen. Syst.* 3: 36-61
- McIntosh, R.P. (1967): An index of diversity and the relation to certain concepts of diversity. *Ecology*, 48, 392-404
- Menares, Brunilda y José I. Sepúlveda (2005): Grupos recurrentes de peces y crustáceos demersales en la zona centro-sur de Chile *Invest. Mar., Valparaíso*, 33(1): 91-100G, r2u0p0o5s recurrentes
- Mulder, B.S., B.R. Noon, T.A. Spies, M.G. Raphael, C.J. Palmer, A.R. Olsen, G.H. Reeves and H.H. Welsh. (1999):. The strategy and design of the effectiveness-monitoring program for the northwest forest plan. USDA Forest Service, Pacific Northwest Station. PNW-GTR-437.
- Nehring K, Puppe C (2002) A theory of diversity. *Econometrica* 70:1155-1198.
- Soriano, silvia v., ana m. Manacorda, nora b. Pierangeli, maria c. Navarro, alejandro l. Giayetto, liliana m. Barbieri, lorena e. Lazzarini, marta c. Minvielle, maría s. Grenovero y juan a. Basualdo (2005): intestinal parasitosis in relation to socioeconomic factors and habitat conditions in children of neuquén, Patagonia, Argentina. *Parasitol Latinoam* 60: 154 – 161
- Stirling, Gray and Brian Wilsey (2001): Empirical Relationships between Species Richness, Evenness, and Proportional Diversity. *The American Naturalist*,. 158, 286–299
- Oehl, Fritz, Ewald Sieverding, Kurt Ineichen, Paul Mäder, Thomas Boller, and Andres Wiemken (2003): Impact of Land Use Intensity on the Species Diversity of Arbuscular Mycorrhizal Fungi in Agroecosystems of Central Europe.
- Patil G. P. and Taile C. (1982) "Diversity as a concept and its measurement," *Journal of the American Statistical Association* 77, 548-567. Pappert, Rebecca A. J. L.
- Hamrick and Lisa A. Donovan (2000): Genetic variation in *Pueraria lobata* (Fabaceae), an introduced, clonal, invasive plant of the southeastern United States. *American Journal of Botany*, 7:1240-1245.
- Pielou, E.C. (1977): *Mathematical Ecology.* Wiley, N.York.
- Pla L. y S. D. Matteucci (2001): Intervalos de confianza bootstrap del índice de biodiversidad de Shannon. *Rev. Fac. Agron. (LUZ)*. 18: 222-234.
- Pothoff, R.R. , M.A. Woodbury and K.I.G. Manton (1992): Equivalent sample size and equivalent degrees of freedom for refinements weights under superpopulation models. *J. Amer. Stat. Ass.* 87, 383-

396.

Qifa Zhang, and R. W. Allard (1986) Sampling variance of the genetic diversity index, *The Journal of Heredity* 77, 54-55

Rennolls, K; Y Laumonier. 2006. A new local estimator of regional species diversity, in terms of 'shadow species', with a case study from Sumatra. *J Tropical Ecology*22:321-329. Renyi, A. 1961. On measures of entropy and information. In: *Proceedings of the Fourth*

Berkeley Symposium on Mathematical Statistics and Probability. University of

California Press, Berkeley, CA. Vol. 1. Pages: 547-561.

Rico-G.;Alejandro, Juan P. Beltrán A.; Adriana Álvarez D.; Eduardo Flórez D. (2005): Diversidad de arañas (Arachnida: Araneae) en el Parque Nacional Natural Isla Gorgona, pacífico colombiano. *Biota Neotrop.* 5,

Royle, J. A. and Dorazio R. M. (2008): *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press, Amsterdam.

Sanders, H.L., 1968. Marine benthic diversity: a comparative study. *Amer. Nat.*, 102: 243–282.

Scheiner, S.M. and J. Gurevitch (eds.) (1993): *Design and analysis of ecological experiments*. Chapman & Hall Publishing, new York, NY

Shannon, C.E. (1948a): A mathematical theory of communication". *Bell System Technical Journal*27: 379–423
Shannon, C.E.

(1948b): A mathematical theory of communication". *Bell System Technical Journal*27: 623–656.

Simberloff, D. (1978). Use of rarefaction and related methods in ecology. In *Biological Data in Water Pollution Assessment: Quantitative and Statistical Analyses*, (Edited by K. L. Dickson, J. Cairns, Jr & R. J. Livingston. West Conshohockeneditors), 150-165. American Society for Testing and Materials. Pennsylvania

Simpson, E. M. (1949): Measurement of diversity. *Nature*. 163, 688.

Smith E.P. (2002), "Ecological Statistics," *Encyclopedia of Environmetrics*, 2, 589-602. (eds Abdel H. El-Shaarawi and Walter W. Piegorsch) John Wiley & Sons. N. YorkSom, R.K. (1996): *Practical Sampling Techniques*. Marcel Dekker, N. York.

Sørensen, T.A. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter*, 5, 1–34.

Scheiner, S.M. and J. Gurevitch (eds.) (1993): *Design and analysis of ecological experiments*. Chapman & Hall Publishing, new York, NY

Shannon, C.E. (1948a): A mathematical theory of communication". *Bell System Technical Journal*27: 379–423
Shannon, C.E.

(1948b): A mathematical theory of communication". *Bell System Technical Journal*27: 623–656.

Simberloff, D. (1978). Use of rarefaction and related methods in ecology. In *Biological Data in Water Pollution Assessment: Quantitative and Statistical Analyses*, (Edited by K. L. Dickson, J. Cairns, Jr & R. J. Livingston. West Conshohockeneditors) , 150-165. American Society for Testing and Materials. Pennsylvania

Simpson, E. H. (1949). Measurement of diversity. *Nature* 163, 688.

Starr, N. (1979): Linear estimation of the probability of discovering a new species. *Ann. Of Statistics*. 7, 644-652.

TortoleroLow, Leonardo José , Dalmiro José Cazorla Perfetti, Pedro Morales Moreno Y María Eugenia

- Acosta Quintero (2008), Prevalencia De Enteroparásitos En Perros Domiciliarios De La Ciudad De La Vela, Estado Falcón, Venezuela. *Revista Científica, FCV-LUZ* , 18, 312 - 319, 2008.
- Valle, A. F. Borges y C. Rincones (2000) Principales malezas en cultivos de caña de azúcar en el municipio Unión del estado de Falcón , Venezuela. *REv.FAc.Agron. (LUZ)* 17, 52-62.
- Valliant, R.; A. H. Dorfman y R. M. Royall.; (2000); *Finite Population Sampling And Inference; A Prediction Approach*; John Wiley, New York.
- Whitford, M. F., Forster, R. F., Beard, C. E., Gong, J. & Teather, R. M. (1998). Phylogenetic analysis of rumen bacteria by comparative sequence analysis of cloned 16S rRNA genes. *Anaerobe* 4, 153-163.
- Whittaker, R. H. (1972). "Evolution and Measurement of Species Diversity". *Taxon*, 21 (2/3), 213-251.
- Yoshioka, Paul M. (2008): Misidentification of the Bray-Curtis similarity index, *Marine Ecology Progress Series, Mar EcolProgSer*, . 368, 309–310.
- Zahl, S. (1977): Jackknifing an index of diversity. *Ecology*, 58, 907-913.
- von Wintzingerode, F., Gossel, U. B. & Stackebrandt, E. (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* 21, 213±229.
- Weitzman M L (1992) On diversity. *Quarterly Journal of Economics* 107:363-405.
- Weitzman M L (1998) The Noah's ark problem. *Econometrica* 66:1279-1298

Capítulo 4

FACTORES RELEVANTES EN LA INCIDENCIA DEL SIDA EN ANDALUCÍA (1992 – 2013)

Yolanda Román-Montoya* y Ana María Lara-Porras*

*Departamento de Estadística e Investigación Operativa. Universidad de Granada. España

Correspondence author: Ana María Lara-Porras

E-mail: alara@ugr.es

Address: Campus de Fuentenueva s/n. Facultad de Ciencias. Universidad de Granada. España

Abstract

AIDS is considered great pandemic of the last few years. This disease is characterized by a number of factors that are associated with their development and that require ongoing review. Data for 5875 AIDS patients are analyzed. These data are from the Andalusian Register of AIDS cases. For each patient, we have information concerning sex, age at diagnosis, transmission category, disease diagnosis, CD4 count. The time period analyzed corresponds to AIDS diagnoses between 1/1/1992 (date on which the application of the dual combination treatments starts) and 09/17/2013.

A descriptive analysis of the measured variables is performed and we use the ANOVA methodology to study the relationships between variables. Finally, we performed a survival analysis for incubation times of HIV virus (time between HIV diagnosis and AIDS diagnosis time). Survival function is estimated using the Kaplan-Meier method and Cox model adjusted proportional hazards.

Key words: AIDS, HIV, HAART, INCUBATION TIME

Resumen

El SIDA, la gran pandemia de los últimos años, se reviste de una serie de características asociadas a su evolución que requieren de una continua revisión. En este trabajo se analizan 5875 datos provenientes del organismo público español Registro Andaluz de Casos de SIDA, que recogen para cada paciente, entre otras variables, la siguiente información: sexo, edad en el momento del diagnóstico, categoría de transmisión, enfermedad del diagnóstico, recuento del CD4. Los datos corresponden a pacientes que contrajeron el virus del SIDA entre 1/1/1992 (fecha en la que se inicia la aplicación de los tratamientos de combinación dual) y 17/09/2013.

Realizamos un análisis descriptivo de las variables medidas y utilizamos la metodología ANOVA para estudiar las relaciones entre las variables. Por último, se lleva a cabo un análisis de supervivencia para los tiempos de incubación del virus VIH (tiempo transcurrido entre el diagnóstico de VIH y el diagnóstico de SIDA). Se estima la función de supervivencia utilizando el método de Kaplan-Meier y se ajusta el modelo de Cox de riesgos proporcionales.

Palabras Claves: VIH, SIDA, TARGA, TIEMPO DE INCUBACIÓN

1. INTRODUCCIÓN

Según el Informe de Vigilancia Epidemiológica del VIH/SIDA, la tasa de incidencia del SIDA en España es de 8.5 por cada 100.000 habitantes, superando la media de la Unión Europea. Esta tasa varía entre las comunidades españolas, destacando la de Andalucía, que es de 1.2 casos por cada 100000 habitantes.

Desde la aparición de los primeros casos de sida en Andalucía a comienzo de la década de los 80 se inicia la vigilancia epidemiológica de esta enfermedad creándose el Registro Andaluz de casos de SIDA, que a su vez forma parte del sistema nacional y europeo de vigilancia del VIH/SIDA desde 1983. Desde que comenzó a funcionar el Registro Andaluz de casos de SIDA hasta septiembre de 2013, se han registrado un total de 12484 casos de SIDA en Andalucía.

A partir de 1993, la vigilancia epidemiológica del SIDA es una actividad incluida en el Sistema de Vigilancia Epidemiológica de Andalucía (SVEA) en todos los niveles de su estructura. El III Plan Andaluz de salud establece objetivos relacionados con la prevención del SIDA que implican su seguimiento epidemiológico, por lo que esta enfermedad se incluye en el programa anual de vigilancia epidemiológica.

Dado que el SIDA se define por la coexistencia de la infección por el virus del VIH más el padecimiento de alguna de las 28 enfermedades indicadoras de SIDA, la fuente principal de información procede de las declaraciones de los médicos del ámbito hospitalario, dónde se producen la mayoría de estos diagnósticos, aunque también se diagnostican casos en atención primaria. Otra institución importante que aporta casos al registro son los centros penitenciarios. Además se hace una búsqueda activa de casos, realizada por técnicos de salud pública, a través de la revisión del CMBD (Conjunto Mínimo Básico de Datos), en el cual se registran las altas hospitalarias de todos los pacientes ingresados en algún hospital andaluz.

Toda esta información se actualiza con datos de mortalidad mediante un cruce de los datos registrados en el Registro Andaluz de casos de SIDA y de los datos disponibles en el Registro Andaluz de Mortalidad.

El Registro andaluz de casos de SIDA ofrece información sobre las características de las personas que la padecen, del lugar donde habitan y de la evolución temporal de esta enfermedad. Hay que tener en cuenta, no obstante, que la incidencia de SIDA es un indicador de los estados avanzados de la infección por VIH en la población y que éste es clave para evaluar la efectividad de los tratamientos y la precocidad del diagnóstico del VIH, que son las intervenciones sanitarias que más impacto tienen en la evolución de la enfermedad tanto en el ámbito individual como en el comunitario; sin embargo, no aporta información sobre la frecuencia de nuevas infecciones por VIH en la población ni de su evolución reciente, pudiéndose dar el hecho que aumenten las nuevas infecciones mientras que los casos de SIDA siguen otra tendencia.

La evolución de esta enfermedad ha tomado diferentes matices desde su origen. Es necesario el seguimiento del transcurso de la enfermedad en los pacientes para identificar las pautas de evolución del virus y de ahí la necesidad de registros concretos y del estudio de los individuos que los integran. Por otra parte, es evidente la necesidad

del estudio del tiempo de incubación de la enfermedad (desde que el individuo se infecta por VIH hasta que desarrolla el SIDA)

Con este trabajo pretendemos cumplir los siguientes objetivos:

- Identificación de ámbitos geográficos y sociales de mayor riesgo y situaciones de especial vigilancia.
- Conocer la distribución de las enfermedades indicadoras de SIDA.
- Analizar la evolución de la epidemia en cuanto a:
 - Vías de transmisión
 - Edad
 - Sexo
 - Enfermedades indicadoras y
 - Mortalidad

2. MATERIALES Y MÉTODO

En el trabajo que presentamos se analizan datos referidos al periodo 1992–2013 procedentes del Registro Andaluz de Casos de SIDA. Se dispone de la información relativa a todos los individuos que han desarrollado SIDA antes de septiembre de 2013.

Hasta esta fecha, la base de datos del Registro Andaluz de Casos de SIDA contiene 12.484 registros, con fecha de alta en la aplicación desde 18 de marzo de 1983 a 17 de septiembre de 2013. Teniendo en cuenta la evolución de la enfermedad y el desarrollo de nuevas terapias de tratamiento, se pueden distinguir cuatro etapas temporales relacionadas con el SIDA:

- 1/9/1975 – 31/12/1986, el virus del VIH aún no era conocido como tal y los pacientes no recibían ningún tratamiento específico
- 1/1/1987 – 31/12/1991, aplicación de la monoterapia
- 1/1/1992 – 31/12/1996, se desarrolla una nueva terapia de combinación dual
- Desde 1/1/1997, se aplica el tratamiento Antirretroviral de Gran Actividad (HAART/TARGA), que consiste en una combinación de tres o más medicamentos anti-VIH) combinado con el establecimiento de programas de metadona.

Basándonos en las modificaciones de las pautas de tratamiento, y en las características particulares de cada periodo, realizaremos nuestro estudio a partir del 1/1/1992; etapa en la que dichas terapias comienzan a tener mejor respuesta en los pacientes afectados de SIDA.

El objetivo de nuestro estudio es la evolución del tiempo transcurrido entre el diagnóstico por la infección del VIH y el desarrollo del SIDA (tiempo de incubación). Tomando como punto de partida el inicio de la aplicación de la terapia dual (1/1/1992), disponemos de un fichero con 5875 registros de los que hemos excluido aquellos registros con diagnósticos de SIDA/VIH anterior a 1/1/1987 y aquellos registros con tiempos de incubación negativo (diagnóstico de VIH posterior al de SIDA).

El Registro Andaluz de casos de SIDA ofrece información sobre las características de las personas que lo padecen, del lugar donde habitan y la evolución temporal de su enfermedad. Concretamente, las variables que lo integran son:

1. Caso de sida (Número de caso o de identificación del paciente)
2. Fecha de alta en la base de datos
3. Fecha de notificación del diagnóstico de SIDA
4. Fecha de nacimiento
5. Sexo
6. Edad del paciente en el momento del diagnóstico
7. Identificación de si el individuo ha fallecido o no (variable dicotómica)
8. Fecha de fallecimiento
9. Código y descripción del país de residencia y del país de origen
10. Código y descripción de la comunidad de residencia
11. Código y descripción de la provincia de residencia
12. Código y descripción del grupo de riesgo que clasifica al individuo en función de los motivos por los que se ha contagiado del virus del VIH. El Registro Andaluz de casos de SIDA distingue las siguientes categorías Usuarios de drogas por vía parenteral (UDVP), transmisión sexual (vía de contagio homosexual o heterosexual), receptores de hemoderivados o transfusiones sanguíneas, transmisión por vía materna, y otras categorías sin especificar.
13. Código y descripción del centro hospitalario
14. Identificación del centro penitenciario en el caso de que el paciente esté o no ingresado en la cárcel
15. Nivel de CD4 en sangre en el momento de diagnóstico del VIH.

Para nuestro análisis hemos seleccionado las siguientes variables:

- Edad del paciente, medida en años y referida al momento de diagnóstico de SIDA.
- Sexo (variable dicotómica con dos valores: hombres y mujeres).
- Enfermedad indicadora de SIDA. Del listado de enfermedades consideradas como enfermedades indicativas de SIDA, se han determinado las siguientes categorías: Tuberculosis; Neumonía por protozoos oportunistas; Candidiasis; Otras.
- Grupos de riesgo: Usuarios de drogas por vía parenteral (UDVP); Transmisión heterosexual; Transmisión homosexual; Otras
- Tratamiento.
 - Terapia Combinación dual: 1/1/1992 – 31/12/1996
 - TARGA y establecimiento de los programas de metadona: 1/1/1997 – 17/9/2013

Tal y como ya hemos indicado anteriormente, el estudio que desarrollamos tiene en cuenta los pacientes registrados desde el 1 de enero de 1992 al 17 de Septiembre de 2013 (aproximadamente 249 meses). A partir del fichero de datos original que consta de 12.484 registros, excluimos 6580 casos diagnósticos de SIDA o VIH anterior a 1/1/1992.

El resto de datos se someten a un proceso de depuración mediante el cual procedemos a la eliminación de todos aquellos registros con información no relevante o errónea. Concretamente, eliminamos:

- 25 individuos, de los restantes, en los que el tiempo de incubación (intervalo de tiempo entre el diagnóstico de VIH y el de SIDA) es un dato negativo o nulo (se

desconoce la fecha de infección por VIH y se asigna como tal la misma que la del diagnóstico de SIDA).

- 9 individuos con fecha de sida posterior a la fecha de muerte (en los 5879 registros actuales)
- 4 individuos sin fecha de edad.

Tras la depuración realizada nuestro estudio se realiza con 5866 registros correspondientes a personas que contrajeron el virus del SIDA entre el 1/1/1992 (fecha en la que se inicia la aplicación de los tratamientos de combinación dual contra el VIH) y el 17/09/2013. Los datos se caracterizan por estar censurados a la derecha ya que en nuestro registro hay sujetos que abandonan el estudio o que no han muerto en la fecha final de nuestro estudio (el evento de interés ocurre después del 17/09/2013).

3. ANÁLISIS DE LOS RESULTADOS

Disponemos de una base de datos formada por 5866 registros de individuos que contrajeron el SIDA ente 1/1/1992 y 17/09/2013, el 93% son españoles, el 3.3% son de procedencia africana y el resto de individuos proceden de diversas nacionalidades tanto europeas como americanas.

Comenzamos nuestro análisis con el análisis descriptivo de las variables numéricas incluidas en el fichero de datos: edad, tiempos de incubación y tiempos de vida con SIDA (en este caso solo consideramos los individuos con fecha de muerte registrada, la inclusión de los individuos censurados se llevará a cabo con un modelo de supervivencia) en función de las modalidades de las variables categóricas incluidas en el estudio.

Del análisis de la tabla 1, se observa que la edad media de los individuos según las categorías de transmisión del VIH oscila entre los 34 y 40 años, el grupo de los contagiados por vía intravenosa son los que presentan una edad media menor, 34.11, frente a los contagiados por contacto heterosexual, 40.43. Se observa también que los tiempos de incubación medios del VIH son inferiores en los individuos contagiados por vía heterosexual (16 meses), observándose unas cifras notablemente mayores en los individuos contagiados por vías intravenosas (31.85 meses). En cuanto a los tiempos de vida con SIDA, se observan también valores mayores para los individuos del grupo UDVP (31.97 meses).

El análisis de estos valores según la enfermedad de diagnóstico de SIDA se resume en la tabla 2. En ella se observa que las edades medias de diagnóstico de SIDA oscilan entre los 35.26 años de los diagnosticados de SIDA por tuberculosis y los 38.31 años del grupo de individuos incluidos en otras categorías. Se observa también que el 50% de los pacientes diagnosticados de SIDA por la tuberculosis presentan también una edad inferior que en el resto de grupos (34 años).

En lo referente a los tiempos de incubación, hemos de destacar que al igual que ocurría con el análisis de la tabla 1, los individuos diagnosticados a edades más tempranas, tienen mayores tiempos de incubación. Se observa que los individuos diagnosticados de SIDA por tuberculosis presentan un tiempo medio de incubación del VIH superior (27.69 meses) sólo superado por los diagnosticados de candidiasis (27.81 meses). En

todos los grupos hemos de destacar la fuerte asimetría de la distribución pues los tiempos medianos en todos los grupos están muy distantes de los valores medios (5 y 2 meses para los grupos anteriormente citados: tuberculosis y candidiasis respectivamente).

Los tiempos de vida con SIDA presentan un comportamiento similar al de los tiempos de incubación.

En la tabla 3, se determinan las características de los pacientes según el tipo de tratamiento recibido. La fecha de diagnóstico determinó la aplicación de uno u otro tratamiento contra el VIH/SIDA. Se observa que las características de los pacientes presentan notables diferencias según el momento del diagnóstico del VIH/SIDA y por tanto, del tratamiento aplicado.

La edad media de los pacientes que recibieron terapia dual (diagnóstico anterior a 1997) es de 34 años, presentando una edad inferior a 32 el 50% de estos pacientes. Los pacientes tratados con TARGA tienen una edad media de 38.6, y el 50% de ellos tiene edad inferior a 37.

Las principales diferencias en estos grupos se observan en los tiempos de incubación del VIH. Mientras que los pacientes tratados con Terapia Dual, su tiempo medio de incubación es de 9.9 meses (sólo un 25% de estos pacientes presentan un tiempo de incubación superior a 17 meses), en el caso de los pacientes diagnosticados a partir de 1997, el tiempo medio de incubación se sitúa en los 34.219 meses, observándose que el 25% de los pacientes supera los 60 meses de incubación del VIH.

En los tiempos medios de vida con SIDA no se observan grandes cambios: 28.8 para los pacientes con terapia dual y 27.56 para los pacientes con TARGA.

Por último la tabla 4 nos muestra los resultados descriptivos de las variables numéricas según la clasificación de niveles del CD4.

El recuento de los valores absolutos de linfocitos CD4 constituye uno de los más importantes marcadores biológicos en el control de las infecciones por VIH/SIDA. Las células CD4 son un tipo de glóbulo blanco muy importante en el sistema inmune; generalmente oscilan entre 600 y 1200. Se sabe que el individuo sero-positivo y con un recuento de CD4 inferior a 200, tiene SIDA independientemente de los síntomas (International Association of Providers of Aids Care). Por tanto, estableceremos tres categorías en el estudio de los niveles de CD4, 0-200, 200-500 y superior a 500.

Se observa que el grupo de pacientes con mayor recuento de células CD4 presenta un valor inferior de la edad media (34.3 años), y el mayor valor de la edad media corresponde al grupo con niveles de CD4 comprendidos entre 0 y 200. En este caso los valores medios y medianos se encuentran muy próximos en los tres grupos de interés. En los tiempos de incubación del VIH, el comportamiento es similar.

		Age (years)	HIV – AIDS (months)	AIDS – death ^(*) (months)
Heterosexual	n	1291	1291	604
	mean	40.43300	16.00155	22.94868
	sd	11.772289	33.01303	34.54411
	25%	31	0	1
	50%	38	1	8
	75%	47	13	27

Homosexual	n	793	793	374
	mean	40.338966	16.41188	24.27540
	sd	11.002903	33.95348	35.90442
	25%	32	0	2
	50%	39	1	10
	75%	47	16	26

UDVP	n	2730	2730	1739
	mean	34.11648	31.85788	31.97815
	sd	6.678421	44.51846	41.84665
	25%	30	0	3
	50%	34	11	15
	75%	38	48	41

Otros	n	1052	1052	290
	mean	37.20913	24.76426	22.05862
	sd	14.305821	47.79668	34.78237
	25%	30	0	1
	50%	37	1	6
	75%	45	24	29

Tabla 1: Análisis descriptivo de las variables numéricas de la base de datos en función de los grupos de riesgo

		Age (years)	HIV – AIDS (months)	AIDS – death (months)
Candidiasis	n	623	623	316
	mean	36.78491	27.81059	33.79114
	sd	10.345881	45.21065	42.34305
	25%	30	0	4
	50%	36	2	15.5
	75%	42	39	42

Neumonía	n	1282	1282	584
	mean	37.25663	21.12090	26.93493
	sd	10.864023	40.33225	37.33244
	25%	30	0	1
	50%	35.5	1	13
	75%	43	23	34

Tuberculosis	n	1945	1945	938
	mean	35.26221	27.69614	38.46055
	sd	8.592162	42.47272	44.65412
	25%	30	0	6
	50%	34	5	21
	75%	40	40	53.75

Others	n	2016	2016	1167
	mean	38.31597	24.00099	19.25364
	sd	11.950637	41.76690	32.09523
	25%	31	0	1
	50%	37	2	7
	75%	44	30	21

Tabla 2: Análisis descriptivo de las variables numéricas de la base de datos en función de la enfermedad de diagnóstico de SIDA

		Age (years)	HIV – AIDS (months)	AIDS – death (months)
Terapia dual	n	2222	2222	1595
	mean	33.97660	9.90234	28.89028
	sd	10.47758	14.07406	40.59210
	25%	28	0	4
	50%	32	1	14
	75%	38	17	31

TARGA	n	3644	3644	1410
	mean	38.69759	34.21954	27.56950
	sd	10.26560	50.15449	37.88183
	25%	32	0	1
	50%	37	2	9
	75%	44	60	40

Tabla 3: Análisis descriptivo de las variables numéricas de la base de datos en función del momento del diagnóstico

		Age (years)	HIV – AIDS (months)	AIDS – death (months)
0 – 200	n	293	293	3
	mean	38.96246	27.57338	66.333
	sd	10.942755	49.92850	80.35131
	25%	32	0	20
	50%	38	0	24
	75%	44	35	91.5

200 – 500	n	107	107	1
	mean	36.96262	60.56075	49
	sd	9.59355	64.00504	--
	25%	30.5	3.5	49
	50%	35	35	49
	75%	42	114.5	49

> 500	n	57	57	0
	mean	34.3333	63.49123	
	sd	8.669185	72.17269	
	25%	28	1	
	50%	33	28	
	75%	40	123	

Missing	n	5409	5409	3001
	mean	36.8248	23.76058	28.22559
	sd	10.6268	40.19221	39.29392
	25%	30	0	2
	50%	35	2	12
	75%	42	32	35

Tabla 4: Análisis descriptivo de las variables numéricas en función de los niveles de CD4 en el momento del diagnóstico

3.1 ANÁLISIS DE LA EDAD DE LOS PACIENTES

Tras el análisis descriptivo realizado, procedemos a analizar la influencia de las variables categóricas en la distribución de los valores de la edad de los pacientes.

Para ello, hemos aplicado un diseño factorial con cinco factores sin considerar interacciones cuyo modelo estadístico es:

$$y_{ijklm} = \mu + \tau_i + \beta_j + \delta_k + \gamma_l + \varphi_m + u_{ijklm}; \quad i, l = 1, 2; \quad j, k, m = 1, 2, 3, 4$$

donde

- y_{ijklm} : Variable respuesta que representa la edad del paciente de sexo i , categoría de transmisión j , enfermedad de diagnóstico k , que recibe el tratamiento l y tiene un número m de niveles de CD4
- μ : Media global, efecto constante común a todos los niveles
- τ_i : Efecto producido por el nivel i -ésimo del factor sexo
- β_j : Efecto producido por el nivel j -ésimo del factor categoría de transmisión
- δ_k : Efecto producido por el nivel k -ésimo del factor enfermedad de diagnóstico
- γ_l : Efecto producido por el nivel l -ésimo del factor tratamientos
- φ_m : Efecto producido por el nivel m -ésimo del factor niveles de CD4
- u_{ijklm} : Error experimental o perturbaciones, variables aleatorias independientes $N(0, \sigma)$.

En la tabla 5 se muestran los resultados del análisis realizado. Se determina la influencia de las variables categóricas en la variable edad. Se observa que todas las variables categóricas estudiadas influyen significativamente en la edad media del paciente. Es decir, existen diferencias significativas en la edad media de los pacientes según la variable sexo, categorías de transmisión, enfermedad de diagnóstico, tipo de tratamiento y nivel de CD4.

	Df	Sum-Sq	Mean-Sq	F-value	Pr(>F)
sexo	1	7706	7706	79.153	<2e-16***
Categorías de transmisión	3	54386	18129	186.161	<2e-16***
Enfermedad de diagnóstico	3	3301	1100	11.303	2.17e-07***
Tratamientos	1	22311	22311	229.183	2e-16***
CD4	3	906	302	3.101	0.0256*
Residuos	5854	569885	97		

Tabla 5: Tabla ANOVA. La variable respuesta: edad media

La tabla ANOVA muestra qué variables son significativas pero no distingue si lo son todas las modalidades. Para estudiar qué modalidades influyen y cuanto influyen vamos a utilizar el contraste de comparaciones múltiples de Tukey que muestra el intervalo de confianza al 95% para la diferencia de edad media entre las categorías de cada una de las variables significativas.

El análisis de comparaciones múltiple que realizamos determina los intervalos de confianza simultáneos construidos por el método de Tukey. Este análisis proporciona un resumen de las comparaciones de cada tratamiento con los restantes. Es decir, aparecen comparadas dos a dos las diferencias en la edad media entre cada par de categorías de las variables consideradas en el modelo. En los resultados que se muestran a continuación obtenemos, los valores **diff**, valor estimado de la diferencia de medias; **lwr** y **upr**, extremos inferior y superior respectivamente de los intervalos de confianza al 95% para cada diferencia.

La tabla 6 ofrece el resultado del análisis de las diferencias por el método de Tukey para la variable categoría de transmisión. Del análisis de los valores concluimos que todos los tratamientos difieren significativamente excepto los grupos de contagio del VIH por transmisión sexual. No existen diferencias significativas en la edad media de los pacientes contagiados por transmisión sexual (p-valor=0.055).

En el resto de categorías si se aprecian diferencias significativas en la edad media de los pacientes. Tal y como ya se observó en el análisis descriptivo, las mayores diferencias provienen del grupo de individuos contagiados por vía intravenosa. En este grupo la edad media es significativamente inferior al de resto de grupos de contagio.

Concluimos que se detectan diferencias significativas en las edades entre los individuos contagiados por vía intravenosa (UDVP) y el resto de categorías, UDVP y los grupos de contagio por transmisión sexual (homosexuales y heterosexuales).

	diff	lwr	upr	p-value
homo-hetero	-1.124925	-2.268885	0.019034	0.0559464
resto-hetero	-3.446449	-4.499566	-2.393332	0.0000000
udvp-hetero	-7.043592	-7.900007	-6.187177	0.0000000
resto-homo	-2.321524	-3.513905	-1.129143	0.0000035
udvp-homo	-5.918667	-6.941488	-4.895845	0.0000000
udvp-resto	-3.597143	-4.517238	-2.677047	0.0000000

Tabla 6: Análisis diferencias significativas en la edad media en función de los niveles de las categorías de transmisión

Si bien este último grupo formado por el contagio por transmisión sexual tiene entre ellos un comportamiento similar en lo relativo a la edad media (p-valor = 0.5594) pero si presentan diferencias significativas con el comportamiento del grupo “otras categorías”.

La figura 1 muestra gráficamente estos resultados. A la vista de la gráfica se observa nuevamente que el grupo de transmisión por contagio homosexual tiene una edad mediana superior a la del resto de grupos.

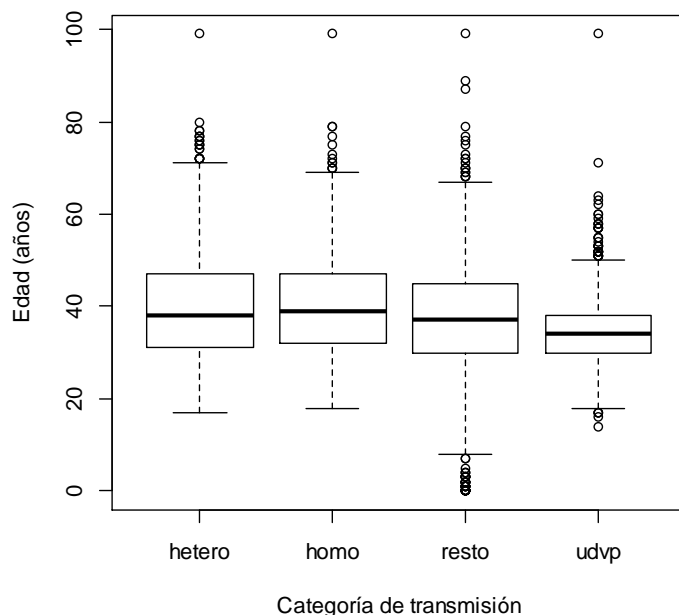


Figura 1: Diagrama de cajas de la edad en función de la categoría de transmisión del VIH

La tabla 7 muestra que las únicas diferencias significativas aparecen con relación a la tuberculosis. La edad media de los pacientes diagnosticados de SIDA por aparición de la tuberculosis difiere de la de los pacientes diagnosticados por neumonía y con la del resto de enfermedades englobadas en el grupo otras. La edad media de estos pacientes (diagnosticados por tuberculosis) se observa en estos dos casos, significativamente inferior. La aparición de la tuberculosis se produce en edades más tempranas que la neumonía o las enfermedades del grupo otras. En el resto de categorías no se observan diferencias significativas.

	diff	lwr	upr	p
neumonia-candidiasis	0.0595	-1.1787	1.2978	0.9993
otras-candidiasis	0.8216	-0.3405	1.9838	0.2655
tuberculosis-candidiasis	-0.958	-2.1256	0.2087	0.1149
otras-neumonia	0.7620	-0.1436	1.6678	0.1340
tuberculosis-neumonia	-1.0180	-1.9301	-0.1058	0.0215
tuberculosis-otras	-1.7801	-2.5859	-0.9742	0.0000001

Tabla 7: Análisis diferencias significativas en la edad media en función de los niveles de enfermedad de diagnóstico

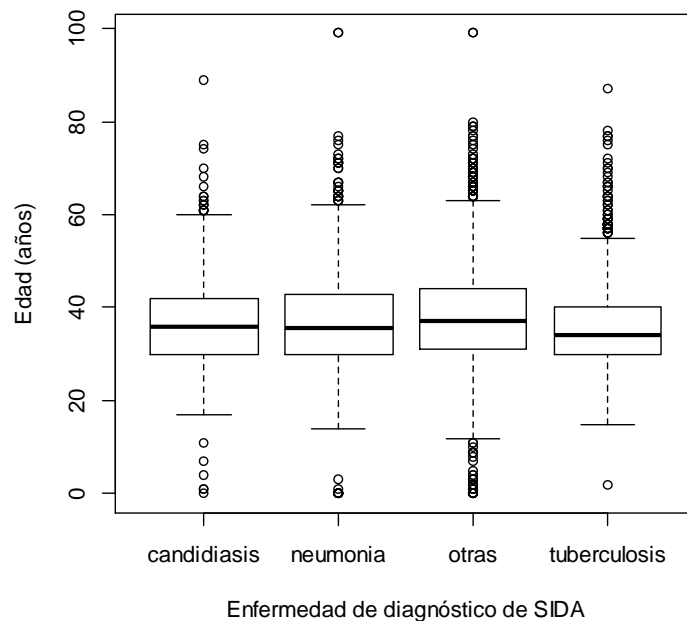


Figura 2: Diagrama de cajas de la edad en función de la enfermedad de diagnóstico de SIDA

La tabla 8 muestra que, con respecto a los niveles de CD4, las diferencias significativas al 90% de confianza con respecto a la edad media de los pacientes sólo se observan en los grupos de niveles extremos [0, 200] y superior a 500.

	diff	lwr	upr	p-value
[0,200] -]500,...]	3.4812	-0.1891	7.1517	0.0703
]200, 500] -]500,...]	1.2515	-2.9061	5.4092	0.8664
Missing -]500,...]	2.1472	-1.2287	5.5231	0.3592
]200-500] - [0, 200]	-2.2297	-5.0937	0.6341	0.1877
Missing - [0, 200]	-1.3340	-2.8549	0.1867	0.1090
Missing -]200, 500]	0.8956	-1.5795	3.3709	0.7888

Tabla 8: Análisis diferencias significativas en la edad media en función de los niveles de CD4

Comprobamos a continuación la idoneidad del modelo propuesto, para ello utilizamos procedimientos gráficos y analíticos.

La figura 3 muestra que se cumple la hipótesis de normalidad de los residuos del modelo. En el gráfico se observan anomalías en las colas de la distribución, lo cual puede venir motivado por la fuerte censura del modelo (el modelo ANOVA es robusto a desviaciones de la normalidad). En la figura 4 comprobamos que se cumple la hipótesis de homocedasticidad e independencia mediante la representación de los valores ajustados por el modelo frente a los residuos obtenidos. Gráfico que no presenta ninguna tendencia sistemática que haga sospechar del incumplimiento de dichas hipótesis.

La independencia de los residuos se ve también confirmada numéricamente a través del contraste de Durbin-Watson que nos permite aceptar la hipótesis de independencia al 95% de confianza (DW = 1.9603, p-value = 0.06119)

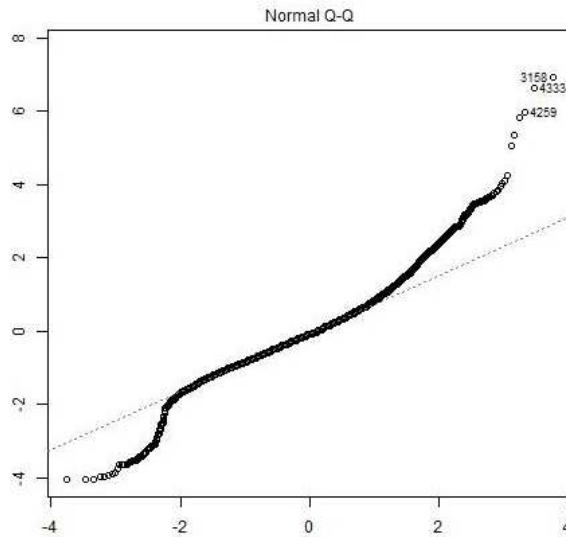


Figura 3: Q-Q plot de los residuos del modelo anova

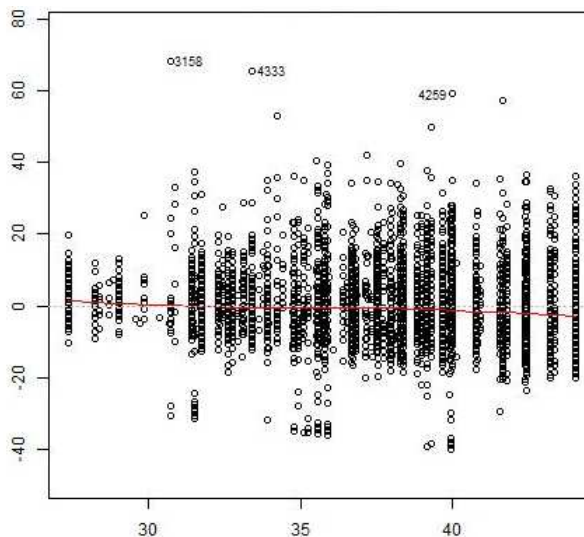


Figura 4: Valores ajustados frente a residuos del modelo

3.2 ANÁLISIS DE LOS TIEMPOS DE INCUBACIÓN

Analizamos a continuación la variable objeto de nuestro estudio: los tiempos de incubación del virus del VIH, es decir, el tiempo transcurrido entre que se detecta la infección por VIH y la aparición de los síntomas de alguna enfermedad indicativa del SIDA.

En primer lugar realizamos un análisis descriptivo de los tiempos medios de incubación (medidos en meses) de los pacientes con SIDA según el año de diagnóstico de SIDA y el sexo. Siguiendo las pautas establecidas en Bueno-Cavanillas et al. (2010) en el estudio realizado sobre 10571 pacientes diagnosticados de SIDA hasta Junio de 2008, se analiza el total de nuestros individuos.

La figura 5 muestra la distribución de los diagnósticos de VIH y SIDA durante el periodo temporal analizado 1992 – 2013. En el caso de los diagnósticos de VIH se

observa un notable descenso en el número de diagnósticos respecto a los primeros años analizados. En el caso de los diagnósticos de SIDA, también se observa este descenso si bien no es tan notable. En ambos casos se mantiene el carácter descendente del número de diagnósticos de VIH y SIDA, lo cual pone de manifiesto la eficiencia de las campañas de prevención.

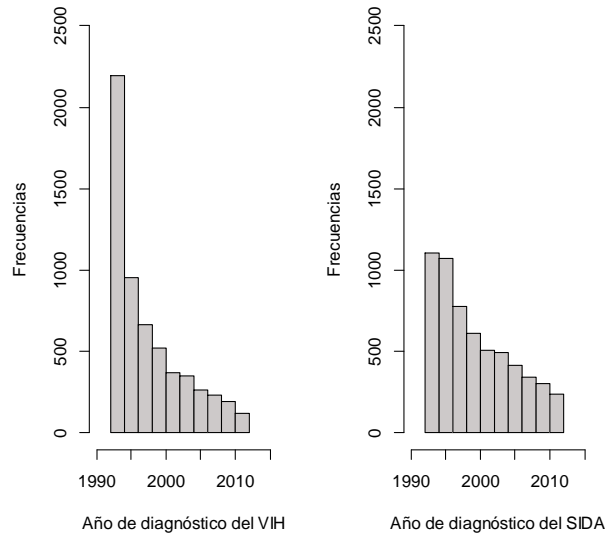


Figura 5: Distribución de las frecuencias de diagnósticos de VIH y SIDA por años.

En la figura 6 se analizan los tiempos de incubación del VIH según el momento de diagnóstico del SIDA, distinguiendo en los individuos según el sexo. En la gráfica de la izquierda se analizan el total de la población y en la de la derecha los individuos contagiados por vía heterosexual; en este punto hemos de señalar que en la base de datos no se contempla la homosexualidad femenina.

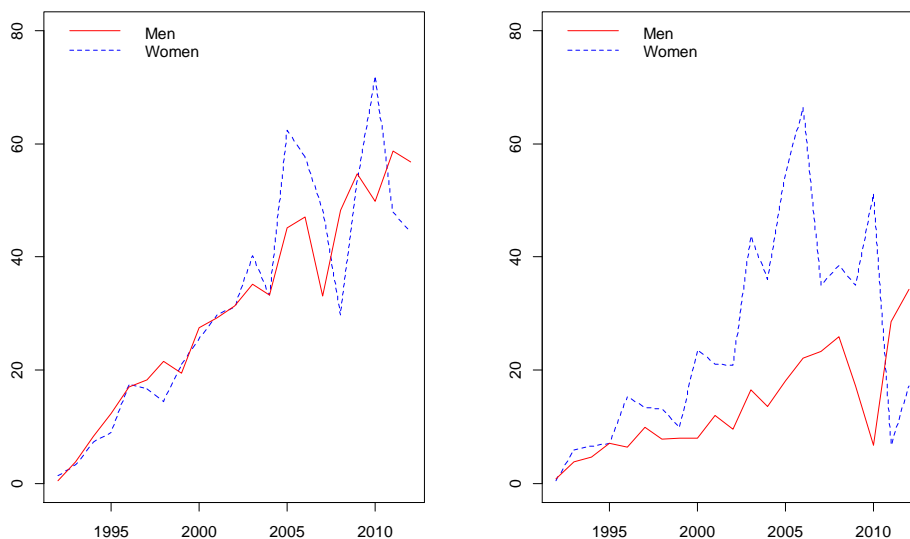


Figura 6: Tiempos medios de incubación según año de diagnóstico de SIDA y sexo.
Izq.: Todos los individuos, dcha: sólo aquellos contagiados por vía heterosexual

En ambas gráficas se observa una tendencia creciente en los tiempos de incubación. En el análisis global de todos los pacientes (Fig. izq), se muestra que los individuos con

diagnóstico de SIDA anterior a 1995 tienen un tiempo medio de incubación (hasta que se desencadenó el SIDA) de unos 10 meses frente a los individuos con diagnóstico de SIDA de 2010 que tienen un tiempo medio de incubación de unos 50 meses (hombres) y en torno a los 70 meses (mujeres). El tiempo de incubación de las mujeres suele ser mayor que el de los hombres exceptuando algunos periodos (1995- 2000) y (2007-2008). Los picos más grandes se observan en 2009 con tiempo medio de incubación de unos 25 meses (mujeres) y 45 meses (hombres). Y en 2011 con tiempos medios de incubación de 70 meses (mujeres) y 45 meses (hombres).

El análisis realizado sólo a los individuos contagiados por vía heterosexual (gráfica de la derecha) muestra un comportamiento muy distinto entre los dos sexos. En los individuos contagiados por vía heterosexual, el tiempo medio de incubación de las mujeres es generalmente superior al de los hombres, siendo la diferencia más notable en el año 2005 en el que el tiempo medio de incubación de las mujeres fue de unos 70 meses mientras que el de los hombres fue de unos 20 meses y en 2010 con tiempos medios de incubación de 50 meses (mujeres) y 5 meses (hombres).

Para concluir podemos decir que al analizar toda la población se observa que no hay grandes diferencias en los tiempos de incubación de la enfermedad según el sexo, sin embargo si consideramos tan solo como vía de transmisión, el contagio heterosexual se observan notables diferencias en la evolución de los tiempos según el sexo del individuo, observándose un incremento considerable del tiempo de incubación de VIH en las mujeres, con mayor relevancia en la etapa del TARGA.

En las figuras 7 y 8 se representa gráficamente el comportamiento de la evolución de los tiempos de incubación de SIDA distinguiendo según categorías de transmisión y enfermedad de diagnóstico respectivamente.

La tendencia creciente de los tiempos de incubación se sigue manifestando en todas las categorías de ambas variables, a excepción del grupo de heterosexuales (fig 7) en el que se observa un descenso a partir del 2007.

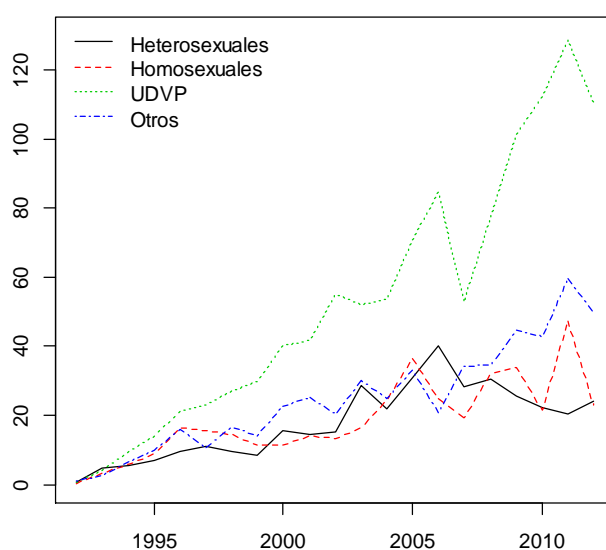


Fig 7: Tiempos medios de incubación según año de diagnóstico de SIDA y categoría de transmisión

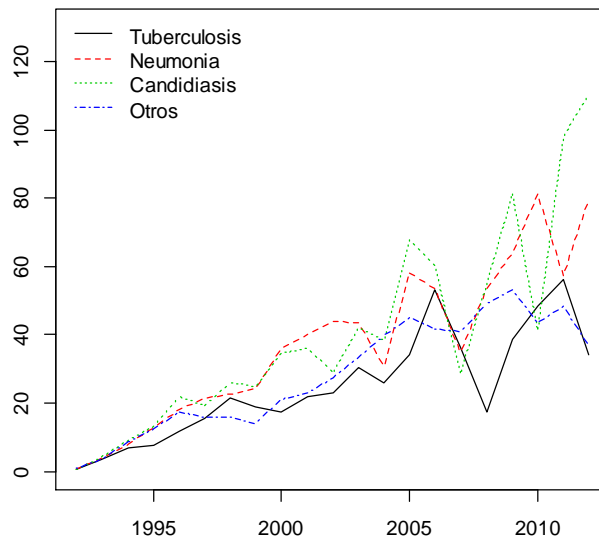


Figura8: Tiempos medios de incubación según año de diagnóstico de SIDA y enfermedad de diagnóstico de SIDA

CURVAS DE SUPERVIVENCIA

Como ya señalamos anteriormente, el análisis de los tiempos de incubación de la infección por VIH requiere de la aplicación de técnicas de análisis específico. El estudio de estos tiempos adquiere una gran importancia, el diagnóstico precoz es uno de los aspectos más relevantes a la hora de tratar a un enfermo de VIH/SIDA. En los últimos años se ha puesto de manifiesto la eficiencia de los tratamientos y con este estudio, queremos destacar los aspectos más influyentes durante los últimos años en la evolución de esta enfermedad.

Para el análisis inferencial de estos tiempos aplicaremos las técnicas de supervivencia. Si consideramos X , la variable *Tiempo de incubación* (tiempo sin SIDA en los pacientes infectados por VIH), la función de supervivencia se define como

$$S(x) = P[X > x] , \quad \forall x \geq 0$$

Como primera etapa en el análisis construiremos las curvas de supervivencia de Kaplan-Meier. Para la determinación de estas curvas se consideran todas las observaciones disponibles en el estudio. Además, este método nos permite evaluar, de modo individual, los posibles factores pronóstico, calcular la supervivencia cada vez que un sujeto presenta el evento y, al igual que otros métodos univariantes, nos permite comparar dos o más curvas de supervivencia a lo largo del tiempo (Armesto, D. 2011).

En la figura 9 se observan las curvas de supervivencia para los tiempos de incubación observadas para el total de los individuos (izquierda) y según sexos (derecha).

La gráfica de la izquierda muestra que 50% de los individuos superan un tiempo de incubación de 10 meses y un 20% superar un tiempo de incubación de 50 meses. Según sexos, se observa (dcha) que el tiempo de incubación es relativamente superior para las mujeres que para los hombres, las mujeres sobreviven más tiempo sin desarrollar ninguna enfermedad indicativa del SIDA, hasta los 140 meses (año 2003) donde el

tiempo es algo inferior en las mujeres o están muy próximos. El 25% de las mujeres no ha desarrollado el SIDA antes de los 50 meses mientras que los hombres está frontera se establece en 40 meses. Sin embargo, tal y como se observa en la tabla 9 de análisis de diferencias entre curvas de supervivencia, el valor mediano es el mismo para ambos sexos, 10 meses. En dicha tabla se observa también que la diferencia entre los tiempos de incubación entre ambos sexos no es significativa para un nivel de significación del 5% (P-valor = 0.0842)

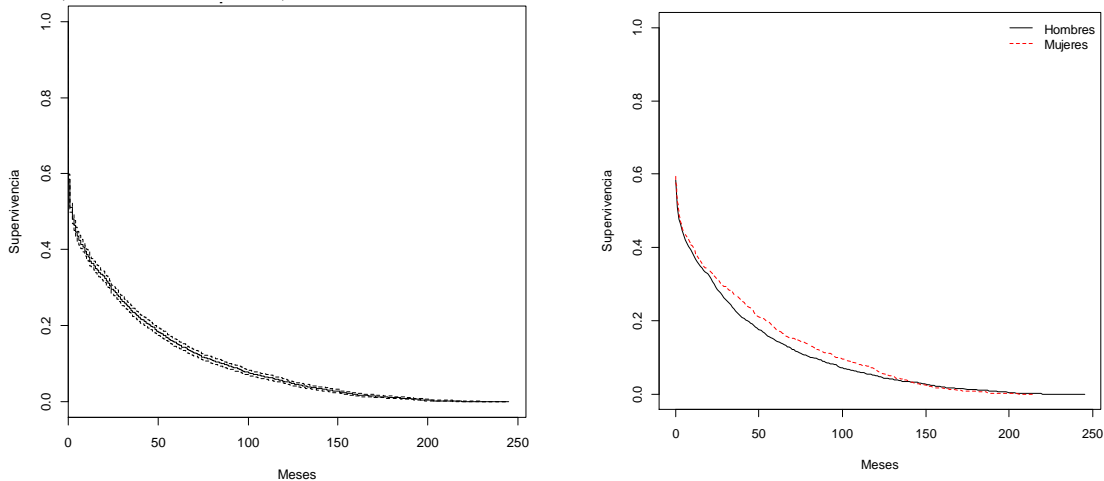


Figura 9: Curvas de Kaplan-Meier de los tiempos de incubación. Izqda: total de individuos. Dcha: según sexos

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
sexo=H	4745	4745	44697	0.496	3.07
sexo=M	1121	1121	1169	1.992	3.07

Chisq= 3 on 1 degrees of freedom, p= 0.0842

Tabla 9: Análisis de diferencias en la curva de supervivencia según sexo

Continuamos el análisis con el estudio según categorías de transmisión y enfermedad en el diagnóstico (figuras 10 y 11).

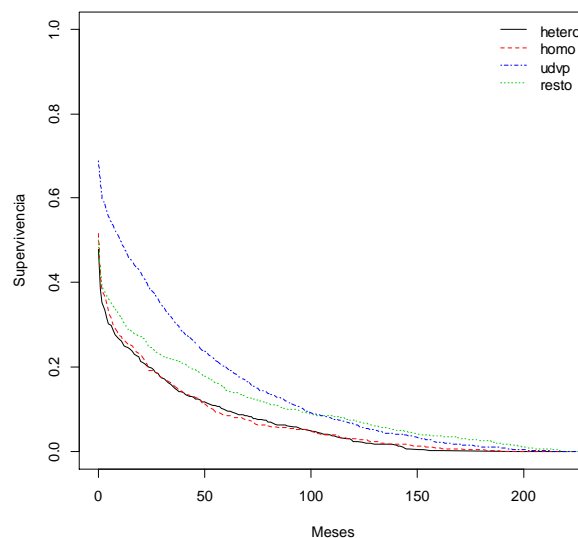


Figura 10: Curva de supervivencia del tiempo de incubación en función de la categoría por transmisión

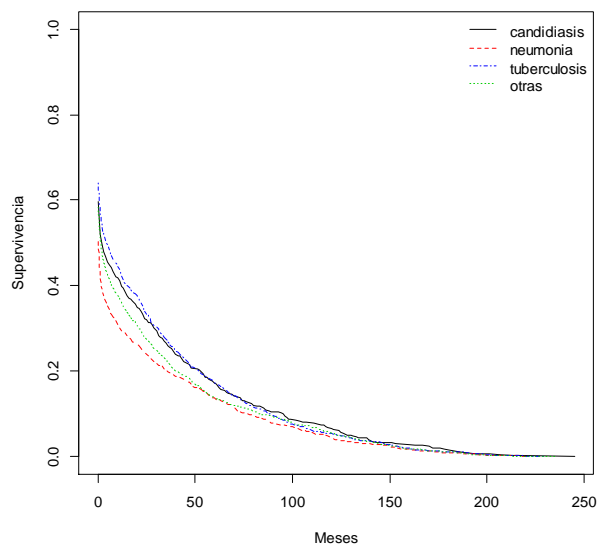


Figura 11: Curva de supervivencia del tiempo de incubación en función de la enfermedad de diagnóstico

La figura 10 muestra que el tiempo de incubación es superior para los individuos contagiados por vía intravenosa (UDVP), hasta los 100 meses (año 1999), a partir de este momento el grupo formado por el resto de los pacientes sobreviven más tiempo sin desarrollar el SIDA. Los pacientes heterosexuales y homosexuales se manifiestan de forma muy similar entre ellos. Observamos que el 50% de los UDVP sobreviven más de 40 meses mientras que los heterosexuales y homosexuales no sobreviven más de 10 meses. A partir de 2008 aunque el tiempo medio de incubación sigue siendo superior para los UDVP, las diferencias entre ellos es cada vez menor. La tabla 10 muestra que la diferencia entre los tiempos de incubación entre las categorías de transmisión se pueden considerar significativa para un nivel de significación del 5% (P-valor <0.000001)

A la vista de la figura 11, se deduce que el tiempo de incubación de la neumonía es el menor mientras que el tiempo de incubación de la tuberculosis es el mayor hasta los 90 meses (año 1998) a partir del cual se mantiene superior el tiempo de incubación de la candidiasis.

La tabla 11 pone de manifiesto que la diferencia entre los tiempos de incubación según las enfermedades de diagnóstico se puede considerar significativa para un nivel de significación del 5% (P-valor=4.11e-06)

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
heterosexual	1291	1291	1030	66.1233	105.10054
homosexual	793	793	649	32.0162	46.7687
udvp	2730	2730	3141	53.7868	146.3694
resto	1052	1052	1046	0.0335	0.0513

Chisq= 198 on 3 degrees of freedom, p= 0

Tabla 10: Análisis de diferencias en la curva de supervivencia según categorías de transmisión

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
candidiasis	623	623	661	2.231	3.09
neumonia	1282	1282	1163	12.107	19.04
tuberculosis	1945	1945	2065	7.024	13.39
otras	2016	2016	1976	0.817	1.53

Chisq= 27.1 on 3 degrees of freedom, p= 4.11e-06

Tabla 11: Análisis de diferencias en la curva de supervivencia según enfermedad en el diagnóstico

El siguiente aspecto a considerar es la evolución según los tratamientos recibidos por los pacientes. Como ya se indicó al principio del estudio, analizamos los individuos diagnosticados de SIDA durante el periodo 1992 – 2013, etapa en la que se han empleado dos tratamientos diferentes. Durante al periodo 1992 – 1996 se aplicaron las terapias duales y a partir del 1996 comienza la era TARGA en el tratamiento del SIDA. Los individuos han sido clasificados en uno u otro grupo según el momento del diagnóstico de la enfermedad (SIDA). Siguiendo los parámetros establecidos en (Román-Montoya et al, 2013), obtenemos las curvas de supervivencia según tratamiento de la figura 12. Tanto gráfica como analíticamente (tabla 12) se observan notables diferencias entre los tiempos de incubación según el tratamiento aplicado, poniendo de manifiesto la efectividad del TARGA.

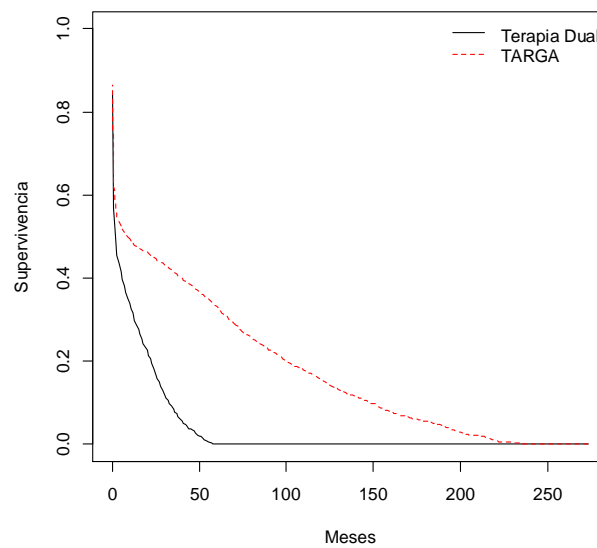


Figura 12: Curva de supervivencia del tiempo de incubación en función del tratamiento recibido

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
Dual	2222	2222	1575	266.2	550
TARGA	3644	3644	4291	97.7	550

Chisq= 550 on 1 degrees of freedom, p= 0

Tabla 12: Análisis de diferencias en la curva de supervivencia según tratamiento

MODELO DE COX

Tras estudiar los factores pronóstico a nivel individual es necesario analizarlos conjuntamente con objeto de evaluar el efecto de las covariables edad, sexo, enfermedad de diagnóstico y categorías de transmisión así como el de posibles interacciones entre ellas.

Como es sabido, el modelo de Cox (Cox, 1972) es un modelo de regresión semiparamétrico que permite determinar el riesgo de ocurrencia de un evento en un momento determinado a partir del análisis conjunto de una serie de variables independientes. Este modelo se basa en la suposición de riesgos proporcionales. En nuestro estudio reviste particular importancia el análisis de las relaciones entre el tiempo de supervivencia y las covariables (proporcionadas por el Registro Andaluz de casos de SIDA) con el objetivo de determinar factores influyentes en el desarrollo de la enfermedad. El modelo de Cox permite, no solo el cálculo de probabilidades, sino que también proporciona una prueba de diagnóstico para medir la relevancia de las covariables incluidas en el estudio.

Los resultados obtenidos de aplicar el modelo de Cox a nuestros datos se resumen en la tabla 13, donde se observa que todas las variables son significativas excepto candidiasis, tuberculosis y el grupo de homosexuales. Analizadas las posibles interacciones entre dichas variables se observa que no son significativas, es decir se comportan igual que el resto (el riesgo de desarrollar SIDA no reviste diferencias significativas cuando se desarrollar por candidiasis o tuberculosis que cuando lo hace por otra enfermedad; análogamente ocurre con el contagio homosexual), eliminando por tanto dichas variables, el modelo resultante queda resumido en la tabla 14.

	coef	exp(coef)	se(coef)	z	Pr(> z)
edad_años	-0.003942	0.996066	0.001435	-2.748	0.0060**
sexo[T.M]	-0.192243	0.825106	0.036592	-5.254	1.49e-07***
candidiasis	-0.086252	0.917363	0.046117	-1.870	0.0614.
neumonia	0.144582	1.155557	0.035912	4.026	5.67e-05***
tuberculosis	-0.046913	0.954170	0.032561	-1.441	0.1496
udvp	-0.336090	0.714559	0.037742	-8.905	<2e-16***
homo	0.081972	1.085426	0.049236	1.665	0.0959.
hetero	0.242494	1.274423	0.042731	5.675	1.39e-08***
tratamiento[T.TARGA]	-0.643980	0.525198	0.030714	-20.967	<2e-16***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tabla 13: Modelo de Cox de los tiempos de incubación

	coef	exp(coef)	se(coef)	z	Pr(> z)
edad_años	-0.003572	0.996435	0.001425	-2.506	0.0122*
sexoM	-0.203731	0.815682	0.035743	-5.700	1.20e-08***
neumonia	0.175702	1.192083	0.031687	5.545	2.94e-08***
udvp	-0.375483	0.686958	0.030814	-12.185	<2e-16***
hetero	0.209479	1.233036	0.037724	5.553	2.81e-08***
tratamientoTARGA	-0.647527	0.523338	0.030601	-21.160	<2e-16***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	exp(coef)	exp(-coef)	lower.95	upper.95
edad_años	0.9964	1.0036	0.9937	0.9992
sexoM	0.8157	1.2260	0.7605	0.8749
neumonia	1.1921	0.8389	1.1203	1.2685
udvp	0.6870	1.4557	0.6467	0.7297
hetero	1.2330	0.8110	1.1452	1.3277
tratamientoTARGA	0.5233	1.9108	0.4929	0.5557

Concordance= 0.607 (se = 0.006)
 Rsquare= 0.117 (max possible= 1)
 Likelihood ratio test= 732.9 on 6 df, p=0
 Wald test = 746 on 6 df, p=0
 Score (logrank) test = 760.1 on 6 df, p=0

Tabla 14: Modelo de Cox de los tiempos de incubación considerando interacciones

Tras la observación de los resultados aportados en la tabla 14, podemos concluir que el modelo es apropiado (todos los contrastes, verosimilitud, Wald y score así lo confirman)

A continuación analizamos la razón de riesgo (probabilidad de que un individuo pueda desarrollar SIDA sabiendo que no lo ha hecho hasta ahora) del tiempo de supervivencia con el virus del VIH hasta el desarrollo del SIDA a partir de la estimación de los efectos de las covariables.

La tabla 14 muestra los resultados de nuestro modelo: coeficientes, exponenciales de los coeficientes (efecto de cada covariable en la función de riesgo), error, valor del estadístico y p-valor asociado al estadístico. Teniendo esto en cuenta, y dado que todos los coeficientes son significativos como ya indicamos anteriormente (p-valor<0.05), podemos concluir lo siguiente:

- Se observa una asociación positiva del tiempo de incubación del VIH respecto a las covariables neumonía y heterosexual, es decir el riesgo de contraer SIDA es mayor en los pacientes que contraigan neumonía y los contagiados de VIH por vía heterosexual. Por el contrario la asociación con el resto de covariables es negativa. Concretamente se observa que:
 - El riesgo de desarrollar SIDA en un mes disminuye un 0.36% conforme aumenta la edad del paciente.
 - El riesgo de desarrollar SIDA es menor en las mujeres en un 18.43%
- Con respecto a las enfermedades de diagnóstico, se observa que es más probable desarrollar SIDA en el grupo de enfermos por neumonía, concretamente un 19.21%. Dado que el resto de enfermedades han resultado no significativas, podemos concluir que el riesgo es el mismo con el resto de grupos de enfermedades.
- En relación a los grupos de riesgo en el momento del contagio de VIH, se observa dos casos relevantes: el riesgo de desarrollar SIDA disminuye en los pacientes contagiados por vía intravenosa en un 31.3% y por el contrario aumenta en los pacientes contagiados por vía heterosexual en un 23.30%.

- Si consideramos los periodos de tratamiento, se observa la eficiencia de la aplicación del TARGA en el tiempo de incubación del VIH. En los pacientes tratados con TARGA, el riesgo de desarrollar SIDA disminuye en un 47.67%.

Una vez establecidos los efectos de cada covariable, procedemos a la validación del modelo ajustado. Se presenta tanto la evaluación global del supuesto de riesgos proporcionales como la evaluación del mismo por cada covariable dentro del modelo. Para ello representaremos los residuos de Schoenfeld y Cox-Snell que nos permiten contrastar las hipótesis de riesgos proporcionales para cada covariable y la idoneidad del modelo respectivamente.

RESIDUOS COX-SNELL

Para cada caso, el residuo de Cox-Snell es la esperanza de que se produzca un evento durante el tiempo de observación

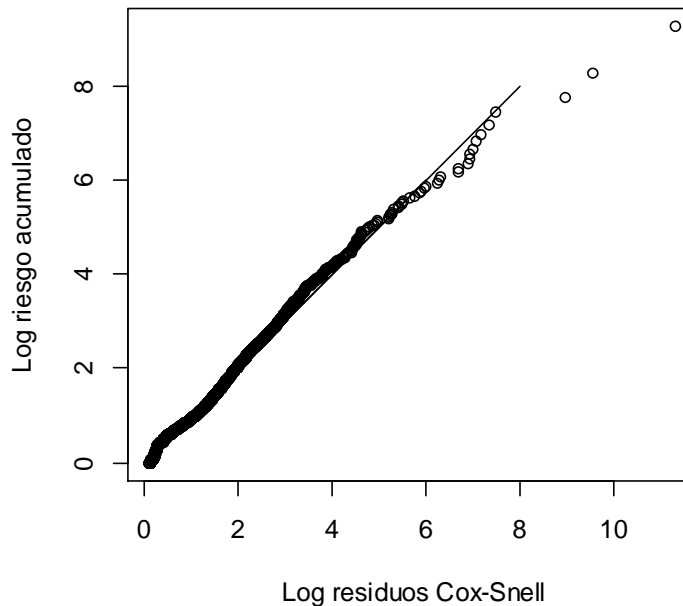


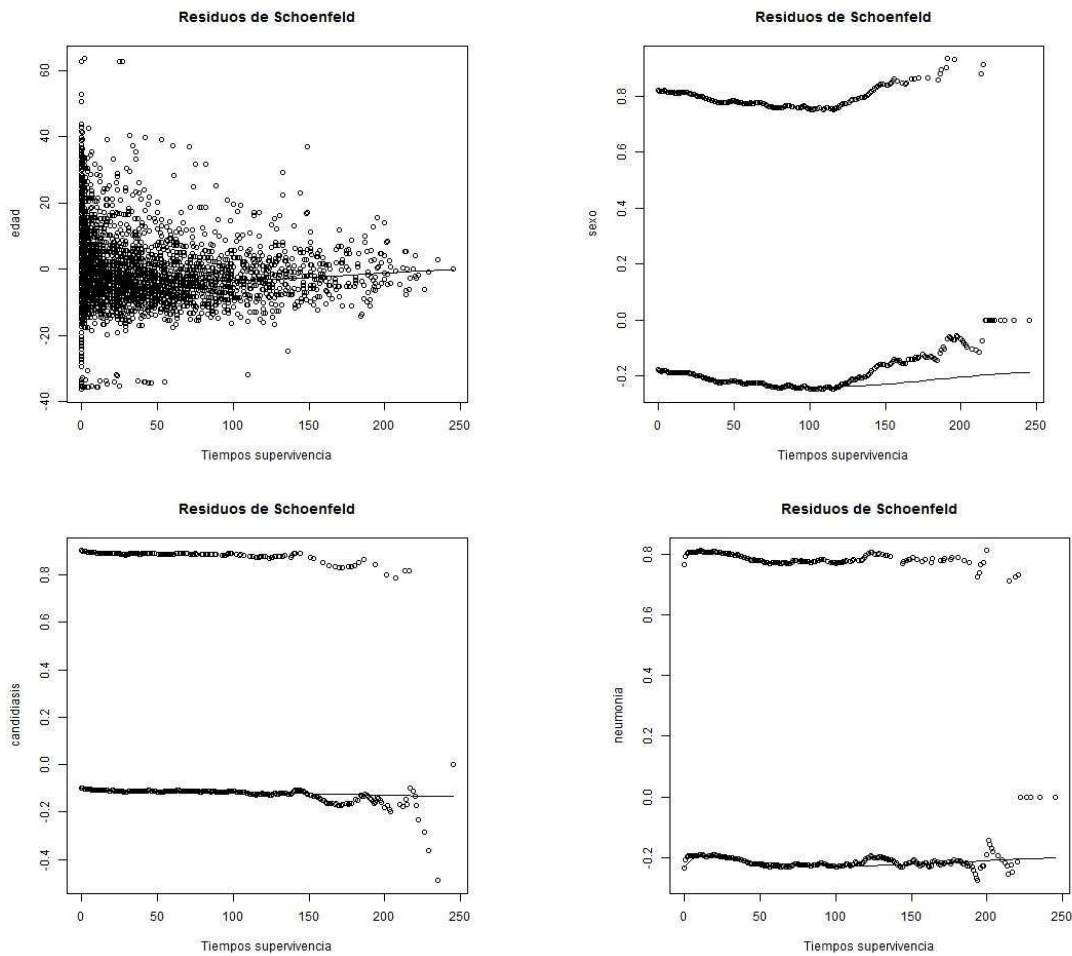
Figura 13: Residuos Cox-Snell

La Figura 13 representa los residuos de Cox-Snell frente a su función de riesgo acumulada. En ella debe observarse una línea recta con pendiente uno que pasa por el origen, que indica que el modelo se ajusta bien. Esta gráfica muestra este comportamiento a excepción de los puntos de la cola.

RESIDUOS DE SCHOENFELD

Los residuos de Schoenfeld son los más efectivos para detectar anomalías para cada una de las covariables que intervienen en el modelo. Se calculan un residuo para cada variable y para cada paciente, en nuestro caso que tenemos un modelo de Cox con seis factores pronósticos se calcularán seis residuos de Schoenfeld por paciente. Estos residuos valen cero para observaciones censuradas o incompletas, por lo que para facilitar su interpretación las salidas de ordenador se suelen presentar sólo para los pacientes fallecidos. Para cada caso no censurado y cada parámetro no redundante del modelo, el residuo de Schoenfeld es la diferencia entre el valor observado del predictor

asociado al parámetro del modelo y el valor esperado del predictor para los casos con el riesgo establecido en el momento del evento observado. Los residuos de Schoenfeld se utilizan para evaluar el supuesto de proporcionalidad. La hipótesis nula especifica que los riesgos son proporcionales, contra la alternativa de que al menos uno no lo es. Gráficamente se manifiesta en curvas paralelas desplazadas una respecto de las otras. La Figura 14 muestra que se cumple el supuesto de proporcionalidad (no hay coeficientes dependientes del tiempo) excepto para la variable edad. En la representación de los residuos de Schoenfeld para el factor pronóstico Edad en función del tiempo de supervivencia, si se cumple la hipótesis de riesgos proporcionales los residuos deberían agruparse de forma aleatoria a ambos lados del valor 0 del eje de ordenadas y la curva ajustada debería aproximarse a una línea recta.



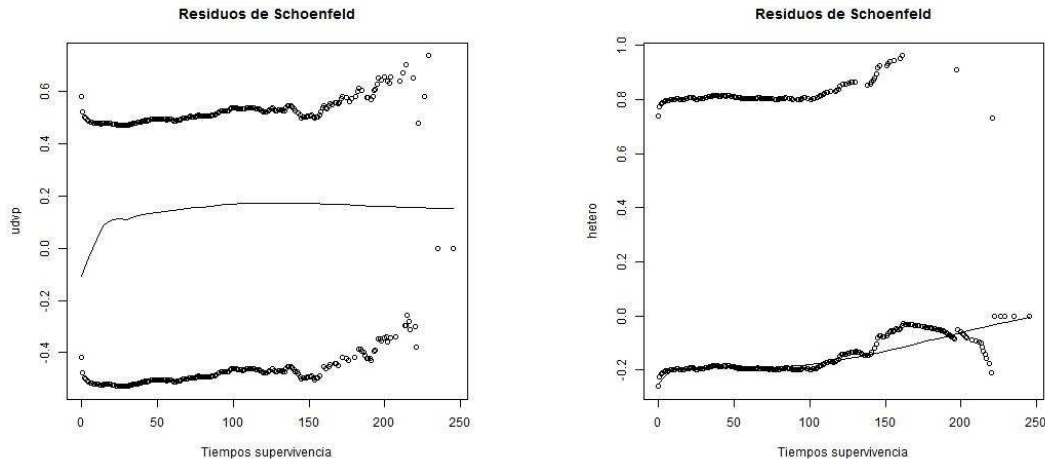


Figura 14: Residuos de Schoenfeld

4. CONCLUSIONES

El estudio de los tiempos de incubación de virus en el desarrollo de enfermedades constituye un aspecto fundamental en el conocimiento de cualquier enfermedad. En el caso del SIDA, el análisis de estos tiempos se reviste de una particular importancia.

A partir de los datos proporcionados por el Registro Andaluz de casos de SIDA analizamos la información relativa al periodo 1992-2013, momento en el que se introdujeron las terapias combinación en el tratamiento de la enfermedad.

Se ha realizado un estudio descriptivo de todas las variables del registro e inferencial sobre la edad media de los pacientes y los tiempos de incubación de los enfermos. En ambos casos se ha analizado la influencia del resto de variables medidas en la base de datos.

Se observan diferencias en la edad media de los pacientes respecto al sexo, entre todas las categorías de los grupos de transmisión excepto los grupos de contagio por vía sexual (homosexuales y heterosexuales) y en los diagnosticados de SIDA por tuberculosis. En cuanto a los niveles observados de CD4 en el momento de diagnóstico de VIH, se observan diferencias en la edad media de los pacientes en los individuos clasificados en las categorías extremas: $[0, 200[$ y nivel de CD4 de 500 o superior.

Para el análisis de los tiempos de incubación hemos recurrido a las técnicas de supervivencia. Del análisis de regresión del modelo de Cox concluimos que a diferencia de lo que ocurría con la edad media, el diagnóstico de SIDA por tuberculosis no es influyente en el riesgo de desarrollar la enfermedad. Por otra parte, en este caso no se observan diferencias en el grupo de contagio por vía homosexual, observándose modificaciones significativas según edad, sexo, desarrollo de neumonía, contagio por vía intravenosa o heterosexual y tratamiento de la enfermedad con TARGA.

REFERENCIAS

- [1] ARMESTO, D. (2011). Análisis de supervivencia. *Rev Electron Biomed* 2: 53-58.
- [2] BUENO-CAVANILLAS, A., GÁMIZ-PÉREZ, L., LARA-POORAS, A. and ROMÁN-MONTOYA, Y. (2010): **Análisis estadístico con datos truncados a la derecha. Estudio de los tiempos de incubación de una muestra de pacientes diagnosticados de SIDA.** En García-Rodríguez, J.F. y Bouza-Herrera, C.N. Investigación aplicada a la salud. Una mirada desde la investigación de operaciones. 117-136. Mexico.
- [3] KLEIN, J. P. and MOESCHBERGER, M. L. (2003): **Survival analysis: techniques for censored and truncated data.** Springer, New York.
- [4] KLEINBAUM DG. (1996). **Statistics in the health sciences: Survival analysis.** New York: Springer_Verlag;
- [5] KOENIG, S., RIVIERE, C., LEGER, P. et al. (2009): High mortality among patients with AIDS who received a diagnosis of tuberculosis in the first 3 months of antiretroviral therapy. *Clinical Infectious Diseases*, 48, 829-831.
- [6] LARA-PORRAS, A., ROMÁN-MONTOYA, Y. y GÁMIZ, M.L. (2009): Regression models with right-truncated data. Applications to the analysis of the incubation time in a sample of AIDS-diagnosed patients. **Proc. of the XIIIth International Conference on Applied Stochastic Models and Data Analysis**, 387–390, Ed: Sakalauskas, L., Skiadas, C. and Zavadskas, E.K.
- [6] PÉREZ-HOYOS, S, [DEL AMO, J.](#), MUGA, R, [DEL ROMERO, J.](#), DE OLALLA, P. G. et al. (2003): Effectiveness of highly active antiretroviral therapy in Spanish cohorts of HIV seroconverters: differences by transmission category. *AIDS* 17, 353-359.
- [8] PEZZOTTI, P., DORRUCI, M., DONISI, A. et al. (2003): Survival, progression to AIDS and immunosuppression in HIV-positive individuals before and after the introduction of the highly active antiretroviral therapy (HAART). *Epidemiol Prev.* 27, 348-355.
- [9] ROMÁN-MONTOYA, Y.; BUENO-CAVANILLAS, A.; LARA-PORRAS, AM. (2013): Evolution of HIV Incubation Times in AIDS Patients. <http://dx.doi.org/10.1080/09540121.2012.748876>. *AIDS Care: Psychological and Socio-medical Aspects of AIDS/HIV.* 25, Issue 8, pages 1051-1061.
- [10] Informe de Vigilancia Epidemiológica del VIH/SIDA en España (2013). Sistema de Información sobre nuevos diagnósticos de VIH Registro Nacional de casos de SIDA. Dirección General de Salud Pública, Calidad e Innovación. Sistemas Autonómicos de Vigilancia Epidemiológica. Centro Nacional de Epidemiología
- [11] International Association of Providers of Aids Care. InfoRed SIDA (2014). URL: http://www.aidsinfont.org/fact_sheets/view/124?lang=spa

Capítulo 5

BETWEEN THE EXPLOITATION OLIGOPSONIC AND FAIR TRADE IN REVERSE LOGISTICS OF SOLID WASTE IN BRAZIL

Mathematical Foundations, Transition Regulatory and Proposal of Policy in Three Steps

Rilton Gonçalo Bonfim Primo¹ and José Félix García Rodríguez²

ABSTRACT

This analysis has for object the process of regulatory transition experienced by the solid waste recycling sector in Brazil, whose market structure is oligopsonic, causing environmental externalities and holding of the base of the reverse logistics of chain represented by scavengers. Making a analysis microeconomic of monopsonic exchange, discusses the mathematical foundations classic of disequilibrium prices in simple and systemic exchange, respectively: 1) adapting the Marxian models of simple and extended reproduction and algebra of analysis input-output 2) introducing equations of trade in differences to study the instability of equilibrium prices. Concludes by proposing the adoption of a public policy to overcome the imbalances and externalities, in three steps: national administration of waste prices by a committee of representatives of the public authorities, civil society and entrepreneurs; direct contract between cooperatives and industries; social reeducation through of innovative systems of stimulus monitored by radio frequency (RFID).

Key-words: Recycling of waste solid. Transition Regulatory. Fair Trade. Foundations Mathematical.

RESUMEN

Este análisis tiene por objeto el proceso de transición regulatoria experimentado por el sector de reciclaje de residuos sólidos en Brasil, cuya estructura de mercado es oligopsónica, provocando externalidades ambientales y exploración de la base de la cadena logística inversa representada por carroñeros. Hace un análisis microeconómico del intercambio monopsónico, discute fundamentos matemáticos clásicos del desequilibrio de precios de intercambio sencillo y sistémico, respectivamente, 1) adaptando los modelos marxianos de reproducción sencilla y ampliada y el álgebra insumo-producto y 2) introduciendo ecuaciones en diferencias para estudiar la inestabilidad de los precios de equilibrio. Concluye proponiendo la adopción de una política para superar los desequilibrios y externalidades, en tres pasos: administración nacional de precios de los residuos por un comité de representantes de los poderes públicos, sociedad civil y empresarios; contrato directo entre cooperativas e industrias; reeducación social a través de innovadores sistemas de estímulos monitorizados por identificación de radiofrecuencia (RFID).

Palabras llave: Reciclaje de Residuos Sólidos. Transición Regulatoria. Comercio Justo. Fundamentos Matemáticos.

¹ Coordinator of Management Advise (Executive Coordination) - *Superintendence of Solidarity Economy* - SESOL of the *State Secretary of Labor, Employment, Income and Sport* – SETRE, by ordinance n° 20101 of the *Government of Bahia-Brazil* (since 10/2009 to 05/2013). Specialist in Production of Economic Information, Social and Geo-Environmental of the *State Secretary of Planning* - SEPLAN of *Government of Bahia-Brazil*, approved in public tender n° 02/2012 of the *Superintendence of Economic and Social Studies of Bahia* - SEI. Accept at 2013, by evaluation of the Scientific Commission, in the PhD Program of Political Economy - *Faculty of Economics, University of Havana-Cuba*.

²Expert of *Secretary de Health of Tabasco.State System of Social Protection in Health*. Teacher in *Universidad Juárez Autónoma de Tabasco (Academic Departure of Economic and Administrative Sciences)*. PhD. in *Philosophical Sciences* by *Center of Studies in Health and Human Welfare of Faculty of Philosophy and History of University of Havana-Cuba*.

1. INTRODUCTION

The prognosticated crisis systemic of employment associated with the configuration of accumulation and concentration intensives in capital constitutes a critical scenario to traditional labor forms, stimulating the alternative creative in associated or autonomous labor, but the worker isolated is a hunted. Thus, it did to reflow of informality to cooperatives a billion people in the world and 30 million in Brazil. It is no coincidence that the *General Assembly of United Nations* (UN) has selected 2012 the *International Year of Cooperatives*, with three main objectives: increase awareness; promote growth; establish appropriate policies. "Through their distinctive focus on values, [said the Secretary-General Ban Ki-Moon] cooperatives have proven themselves a resilient and viable business model that can prosper even during difficult times." The systemic dereliction to the investment in 2008, with major consequences, has restructured this global tendency, closing old entrance and opening innovative windows. In others words, the "socioeconomic crisis that, starting from Wall Street, struck the world, offers to Brazil and several countries the opportunity to introduce changes in its development strategy", signals, from *Ecole de Hautes Etudes en Sciences Sociales* (EHESS) of Paris, Ignacy Sachs, for who, "in the context of a market economy, may wish to broaden the scope of social and solidarity-based economy, whereas this is not governed by the principle of individual appropriation of profits" (Sachs, 2009, p. 139-140). However, this process of social reintegration is not linear, and is not always symmetric. If the dialectic relation between center and periphery in the dynamics of 'metropolitan areas' of Brazil is not new or neutral, the bartering between suppliers and urban bargain hunter of recyclable materials are still less, dating back to the reverse logistics of pauperism in Brazil Empire.³ Office-historically dispersed around sinuous prices, the *Brazilian Institute of Geography and Statistics* - IBGE points that, in Brazil, still yesterday,

[...]recycling is, for the most part, the result of the activity of scavengers (autonomous or organized in cooperatives), and is not a result of a compromise more profound and widespread population and the authorities with the process of separation and selective collection of garbage. In this context, the percentage of recycling of materials fluctuates greatly as a result of fluctuations in the prices of industrial raw materials and the level of employment. (Vilanova Neta, 2011, p. 4).

In fact, "the scavengers are the basis of the productive chain of recycling - as it has been estimated that 90% of all the recycled material in Brazil is recovered by the hands of these agents" (Freitas, Fonseca, 2012, p. 9). One should mention, in competitive terms, "only 10% of the gatherers are linked to a collective organization. The vast majority still operates in isolation, in precarious conditions and without the possibility of access to best selling prices of recyclables." (Id., p. 60). Jumps to the eyes of the issue of the relative weakness of recyclers at a market highly centered dictating prices and condi-

³ Among the self-employed, then, include, e.g., "the 'recyclers', divided into two lineages clearly different - of that surreptitiously collect rags clean and the dirty rags; the 'wastes'; the 'cavaqueiros', which remove the mountains of garbage in search of objects and materials bought; the 'plumbers', pickers of remnants of lead; the 'hunters of cats', purchased by restaurants where they were sold as rabbits; the 'collectors of boots and shoes'; the 'catch-labels' and postage stampers, who were looking for labels of imported items and stamps of fine cigars to sell them to counterfeiters; the 'mousers', which they bought the rats dead or alive to private individuals to resell them to the Board of Directors of Health; the 'readers of hand', the 'tattoos', the 'hawkers' of prayers and Literatures of twine and the composers of 'modinhas'". (Sevcenko, 1999, p. 60).

tions. It is not a regional exception but a conflict, indigence against opulence, in very poor regions.

The major benefit of the entire process of recycling garbage in Brazil and in the world are the industries, precisely because they are highly concentrated, representing a model of oligopsony, when a small number of companies consumes the recyclable materials and imposes the conditions and prices to the gatherers and cooperatives, making them the hostages of the exploitation of the formal economy on the informal. This is the sad situation in which these agents to working with garbage in our country. With cooperatives/associations or with collectors autonomous, the exploitation happens in all levels of the sequence productive and reproductive of recycling. Under these conditions, the recycling, in itself, does not represent a cost-effective alternative and much less environmental; only smothered shortly the social pressures on the unemployment of the excluded and provides a gain by industries, through the reduction of their costs; and these, using the scrap, the great "lords of rubbish", control the market for recycled products. (Magera, 2008, p. 17).

In addition to the reverse logistics required by law, such as the pesticides, their waste and packaging, batteries, tires, fluorescent lamps, electronic products and their components, etc., "other waste can be objects of the string of reverse logistics, for example, medicines and packaging in general" (Fernandez, 2012, p. 9), these in particular. In the field of recycling urban waste economically and environmentally more relevant in Brazil are containers and other objects of glass, paper/pulp, plastics, aluminum and steel. The *Scavenger of Recyclable Material* since 2002 is in *Brazilian Code of Occupancies* (CBO), but is very little recognized. Nowhere the scenario is more delicate formal interpretation, in the case of economic actors more invisible, without steady income, outsiders of formal jobs, at least when they are not in any way meeting. Rigorously, nor would it be possible to calculate a *rate of exploitation* of this workers⁴, because it is calculate on the salary which, in common, they do not have⁵.

The scavenger also would be associated with the term "exploitation. According to the literature, they would be operated by other sectors more articulated chain (Buenrostro, Bocco, 2003; Gonçalves, 2003; Kaseva, Gupta, 1996; Magera, 2003), due to the very structure of the market for recycling and the sectors that dominate - intermediaries and industry -, being common training of oligopsonys informal, that lead to a control of the price paid by the material. (Carmo, 2011, p. 9).

⁴ "In response to the virtual loss of source of income, the path already paved by the gatherers is the collective organization of category. The gain immediate to the scavenger organized is with respect to the terms and conditions of work that is in cooperatives, as regular day as personal protective equipment and sanitary conditions more suitable for the performance of its activities. However, the employment relationships are still very weak in most of the cooperatives and associations - only in a minority of cases, these organizations have legal record and are able to observe the entire labor legislation and tax, and in most cases do not have equipment and sufficient knowledge to ensure economic efficiency. The elaboration of public policies geared to the gatherers will require a more in-depth knowledge of the current situation of organizations of collective work." (Freitas, Fonseca, 2012, p. 17).

⁵ The rate of exploitation (Ω) would be calculated at the same time that $\Omega = \frac{\sum_{i=1}^n S}{(R_t - D)}$ in which S are the gross wage average of n workers, R_t total revenues and D the spending of period, given by (*Fixed Costs + Variable Costs*).

The discussion of *Fair Trade* today is renewed with technical and empirical reviews:

The issue of marketing is quite complex and involves several aspects that deserve special attention, since the limit for the improvement of the conditions of marketing cooperatives. Even if they have gained some power of negotiation of prices, the recycling market, as explained by Gonçalves (2003, p. 143), is an oligopsony, i.e. there is only a small number of purchasers: "Without competition, a product such as the recyclable material separated from the source (or not) is sold at a price that the oligopsony want to pay, which keeps the cost of reverse flow extremely cheap." The power of industry is evident in the analysis of Calderoni (2003) that attaches to this segment of the value chain up to 75% of total gains possible provided by recycling (Grimberg, Tuzel, Goldfarb, 2004, p. 18-19).

The researcher R. Smith, in his *Unequal Trade and Industrialization in the Northeast*, explained that, contemporaneously, involucre into the free trade, "the unequal exchange is focused as the exchange of quantities of work (or value) not equivalent", or by economics "that admit the non-equivalence in trade as a result of a mismatch in transitional competition, since it does not lose the referential of equilibrium", or for "those who admit to not transience of remoteness in market prices in relation to the prices of production. Is the oligopolistic capitalism that determines the permanent nature of this remoteness" (Smith, 1985, p. 10). Precisely this is the scenario of concentrated sector.

In the cycle of recyclable materials, recycling companies located in upper circuit using a technology of relatively high level of expertise, usually capital intensive from the advanced countries. Thus, the activities of the upper circuit tend to concentrate power and control the cycle as a whole. It is important to remember that the market for recyclable if configures, on the demand side, as an oligopsony - few companies have the ability to purchase certain goods -, which further demonstrates the concentration exerted by upper circuit. This way, the materials that are collected by the workers of the lower circuit (which make up a situation of "perfect competition" on the supply side) are separated and then sent to the companies. In this process the material will having its value increased, as it move through the cycle in the direction of the upper circuit (Dagnino, Dagnino, 2010, p. 71).

In fact, the markets for solid waste in Brazil are structured as oligopsonys within which proliferates a floating number of middlemen who sign frequent price agreements (and other conditions of purchase) that not only generates inefficiency in the process of allocation of resources, with reductions in quantities recycled and the level of welfare of society; in particular, they degenerate in demeaning the work values of the gatherers, whose income/product is transferred, in a staggered form, to the links above the value chain of the industry. Nationally, the demand direct focuses on traders (middlemen), which are more than half of the supply of waste, followed by industry (addressed to direct and indirect end), with less than 20% (Brazil, 2008). The professor of the *Graduate Program at the School of Management and Business, University of Rio Grande*, Maria do Carmo, portrayed the local scenery⁶ in a convergent form to the Para-

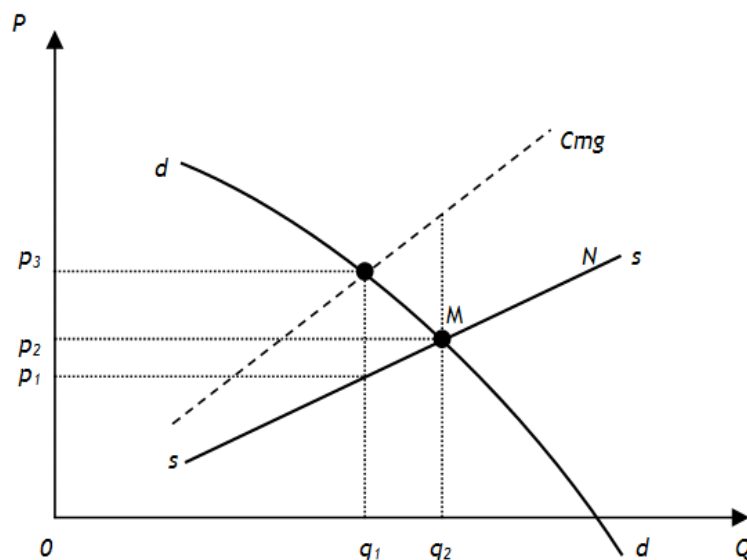
⁶ "Unlike the gatherers, buyers would be well articulated, because the paper trading in the city of Rio de Janeiro is marked by the union of three large middlemen. This union gave rise to Recycling Center of Rio de Janeiro (CRR), - considered to be the main buyer of material collected by collectors in the city, an oligopsony." (Carmo, 2009, p. 62).

ná⁷. Among other urban concentrations, the biggest example is São Paulo with your scenario of Piracicaba, among other several examples sub-regional:

The marketing of the material collected, separated and pressed by the Cooperative is made by own recyclers from the realization of a register of buyers. The so-called 'scrap' acquires the materials collected by recyclers and determines prices by reference to the selling prices (prices at which the materials will be sold to companies that recycle - companies end). Already in the case of the relationship between the scrap with the companies end are the latest that determine prices, setting in some cases a situation of oligopsony. (Martins, Peres, 2006, p. 3).

2. MICROANALITICAL APPROACH, CLASSICAL THEORY AND UNEQUAL TRADE

Clearing the possible lines of regulatory action on markets above mentioned, will make a brief microanalysis theirs, whereas the practical effects of the agreements between the oligopsonists approach these environments to the standard monopsonic, of which, for clarity and simplicity of exposure, will be as such treaty, not compromising the consistence of reflection.



Graph 1: *Dynamic of Monopsonics Markets*
Source: Elaboration of the author.

The functional results of "adjustment of minimum prices" are often mentioned among the measures to mitigate, when not to neutralize the adverse effects of power monopsonic, as shown in *Graph 1*, where are the possible prices (P) and quantities (Q), the supply curve (ss) of the gatherers, the demand (dd) monopsony and marginal cost (Cmg). Reign that the plaintiff, "seeking to maximize the profit, restricts the amount of resource used

⁷ "In the four cities (Paraná) researched, there is no recycling industries installed. All the recyclable material collected is pressed and sent to companies that recycle cardboard, paper, aluminum, glass and plastic. In the case of Paraná, are five the recyclers of paper registered in Cempre (Corporate Commitment to Recycling). If this number is compared to the observation that picking recyclable is a reality in more than 200 Parana municipalities that are related to the Project 'Parana Environmental', it goes to a characterization of these recyclers as forming a oligopsony" (Bosi, 2008, p. 108).

(from q_2 to q_1) and paid unit price less than the marginal revenue of corresponding product " (Leftwich, 1974, p. 323), i.e., p_1 not p_3 .

Derived from two restrictions of economic system of capitalist accumulation that all monopsonic markets behave this way, which are: (A) the contender with power monopsonic is encouraged to purchase up to the point of intersection of the curves of demand and marginal cost (ordered pair p_3, q_1); (B) the curve of marginal cost (Cmg) is higher than that of supply (ss), because they must pay, at the price of the extra unit (marginal), all units purchased (Salvatore, 1984, p. 398-429). Given an initiative that, in this *Graph 1*, takes p_2 to a minimum price of the residue, the amount and the equilibrium price may be M , neutralizing the power of monopsony and returning to the gatherers to rent/product transferred to monopsony. Other, eliminated the demeaning of prices, will be drawn on the environment waste not collected ($q_2 - q_1$). Of course, in any case will fit attempts to price adjustment that neutralize the monopsony, being that, for every moment in a real situation, "such precision can be reached or not. However, any price between p_1 and p_2 would also counter equilibrium the monopsony. The closer to p_2 is the set price, closer to elimination of exploitation" (Leftwich, 1974, p. 329), as well as of environmental externalities, when price/value = 1.

Unmake the apparent paradox mathematical of postulates $D_t = D_{t+1} > D_t$, when, for example, the value of a ton of waste, in line ascending of the value chain, from the scavenger loose to industrial consumer, is bought by a quantity of money D the initial moment t and resold in moments $t + 1, t + 2, \dots, t + n$, for a sum of money greater than the original, or is that there has been adding value to the residue in the form of human labor, or that is the retroactive effect of the power of monopsony, or for both of these reasons. The law of accumulation private and concentration, finally, responds by fetishized capital productivity that co-transforms the residue in gross condition. Thus, the *process of production* of aluminum, paper, glass or plastic recycled, generating a total value end higher than the initial can be highest simplified in this chain of three equations + one inequality:

$$D_t = [(M_t^1 + M_t^2 + \dots + M_t^n) + (L_t^1 + L_t^2 + \dots + L_t^w)] = (M_{t+1}^{n+1} + M_{t+1}^{n+2} \dots + M_{t+1}^m) = D_{t+1} > D_t$$

Or, more concisely,

$$D_t = \sum_{i=1}^n M_t + \sum_{i=1}^w L_t = \sum_{i=n+1}^m M_{t+1} = D_{t+1} > D_t$$

where D_t is the money in advance during the period t necessary to the production line of final goods, M_t^1 is the value of the first of these n goods acquired with productive determined purpose, including the inputs from the recyclers, L_t^1 is the quantity of labor socially necessary type 1 of other w types used in this production, including recycling, M_{t+1}^{n+1} is the value of the first m goods produced at time $t + 1$, D_{t+1} is the sum of the value of these goods, greater than the sum D_t advanced in *raw materials, inputs and wages* (v), being

$$D_{t+1} > D_t \Leftrightarrow \exists L_t^\alpha > v_t^\alpha$$

where v_t^α is the remuneration of L type α at time t , and $1 \leq \alpha \leq w$. Stay explained that $D_t = D_{t+1} > D_t$ because $D_t = \sum_{i=1}^n M_t + \sum_{i=1}^w L_t = \sum_{i=n+1}^m M_{t+1}$. But, thing dis-

tinct occurs when the profit is only commercial, when the plaintiff makes a profit without adding value to the residue, but because buying it below the real value. We not intend in here to discuss how, in the sphere of production, can occur $L_t^\alpha > v_t^\alpha$ and how $\sum_{i=1}^n M_t$ can be retransmitted full to $\sum_{n=1}^m M_{t+1}$ without generating any part of $\Delta D = D_{t+1} - D_t = D' - D$, since that

$$\frac{M_t^\beta}{P_t^\beta} = 1 \forall \beta$$

where P_t^β is the price of the goods β at time t where $1 \leq \beta \leq n$. This last condition, quotient unit, however, it is almost never met in concentrated capitalism. This result fruitful is often neglected. In Marxian terminology, if the *constant capital* is purchased below its value, it becomes variable (generator of ΔD) in the same way that the advance in wages, which alters the original theoretical formulation for

$$D_{t+1} > D_t \Leftrightarrow \exists L_t^\alpha > v_t^\alpha \text{ and/or } \exists M_t^\beta > P_t^\beta$$

although, considering the economy as a whole $\sum \Delta D = \sum_{i=1}^w L_t - \sum_{i=1}^w v_t$, always, because when $\exists M_t^\beta > P_t^\beta$ in a given production process, in a symmetric form, $\exists M_t^\beta < P_t^\beta$ in at least one production process, nullifying the *departures folded* where macroeconomic Σ losses - Σ earnings = 0. It is, in this respect, the case of when the plaintiffs of solid residues influence the prices, with losses related to the sellers. The reverse in the offer. Such inferences will be extrapolated to n markets in *Appendix A*, where it demonstrated the systemic consistency of this argument line.

More easily accepted when the analysis of the imperfections of the market, this result becomes controversial when introduced in the productive processes, but do not assume or imply in hypothesis of, this rather controversial, 'productivity of capital', which postulates that all or at least a fraction of ΔD is explainable by the consumption of goods β even when $P_t^\beta - M_t^\beta = 0 \forall \beta$. Insists the neoclassical theory that is all, because every penny of value ΔD in the form of profit (interest etc.), is pocketed by capitalist, fact that indicates only the final destination of the surplus, not its origin. Against this hypothesis of an object, extrinsic to the human being, with the ability to generate new value, D. Ricardo had pointed in his *On the Principles of Political Economy and Taxation* (London, 1821, pp. 336-337), which already

[...] Adam Smith not underestimate in any place the services that natural agents and the machinery in render but distinguishes very precisely the nature of the value they add to goods [...] as they carry out their work of charge free, the aid that we provide nothing adds to the value of an exchange. (Ricardo apud Marx, 1996b, p. 40. 22n).

On this note, K. Marx said: "Of course the observation of Ricardo is correct against J.-B. Say, that can imagine that the machines provide the 'service' to create value, which is part of the 'profit'." (Marx, 1996b, p. 40. 22n). Written between the years of 1861 to 1863 to compose the *Book 4 of The Capital*, Part 3 of the so-called *Economic Manuscripts* (or *Grundrisse*), it elucidates the apparent *productivity of capital* in these terms:

Since the live work - with the exchange between capital and worker - is incorporate in the capital and appears as your activity, since the beginning of the work process, all productive forces of social work pass to play the role of productive forces of the capital, in the same way that the social form of labor appears in money as the property of a

thing. Thus, the productive power of social work and its particular forms if present then the quality of productive forces and forms of capital, labor materialized, the material conditions (objective) of the work - which, in this independently, in face of live work, embody in capitalist. Here, once again, the inverse relationship that, in dealing with the money, we call fetishism. [...]. In its simplicity, this relationship is already a perversion, the personification of the thing, and reifying the person; because what distinguishes this form of all of the above is that the capitalist dominates the worker does not by virtue of a personal attribute, but only as long as it is "capital"; this might just work materialized $[\sum_{i=1}^n M_t]$ on the lives $[\sum_{i=1}^w L_t]$, the product of the worker on the worker. But the relationship is even more complicated and look more enigmatic because, with the development of the mode of production specifically capitalist, are opposed to the worker and against in the role of "capital", in addition to these things immediately materials - they all work products [...] - the forms of work socially developed, cooperation, manufacturing (a form of division of labor), plant (form of social work organized on the basis material in machinery), representing forms of development of the capital, and, therefore, the productive forces of the work developed, from these forms of social work, as a result also the science, and the natural forces, appear as productive forces of capital. (Marx, 1987, pp. 384-385).

This debate was resumed in *Cambridge Controversy* (Cf. Robinson, 1953; Samuelson, 1962; Pasinetti, 1969) on the *theory of capital*, debate between the results of which it is said that if ΔD it is the owner of D_t , regardless of $P_t^\beta = M_t^\beta$ or $P_t^\beta \geq M_t^\beta$, this occurrence is not founded because the economic science see goods are born, by chance, of other goods, but "necessarily to see the capital $[M_t^\beta]$ as a social object - and not as a mere factor of production, a physical object" (Teixeira, 2004, p. 1), so that, at time $t+1$, having origin exclusively or not in v_t^α , for any specific industry, "the difference $[\Delta D]$ if forwards it to the owners of the property, because they monopolize the means of production $(\sum_{i=1}^n M_t)$." (Harris, 1974, p. 563). The relationship of ownership is fetishized in *productivity* of things (capital). The latter was mathematically measured, zero. The supply/demand equilibrium is usually different of unitary fraction price/value. The mathematical bases of analysis of the processes of formation of disequilibrium price composes the *Appendix B*, that only introduce, does not exhaust the subject.

3. REGULATORY AND TECHNOLOGICAL OPPORTUNITIES OF CHALLENGERS

In 2010 the *Ministry of Environment* requested the highest scientific organization of advice political-economic of the Presidency of the Republic, the *Institute of Economic Research Applied* - Ipea, an estimate of how the Brazil loses with the gap between the amount of recyclable waste and effectively recycled, and the Ipea has reached the figure of R\$ 8 billion annually (approximate US\$ 3,48 billion). In fact, even though 60% of the municipalities in the country is already in possession of some initiative of selective collection, the quantity which in fact returns to the productive chain does not reach 2% (Agência Brasil, 2013). This framework is changing at an increasing rate, and is expected to become deeply in ten years, or the rationalization of logistics will be much more costly and emergency, calculates the *Brazilian Association of Cleaning Companies Public and Special Waste* (Abrelpe). Already surpass the positive inflection point. The present proposal, were found the demand to offer, can elevate the volume of solid waste sent to the reverse logistics industry.

More broadly, in other research compiled by Ipea, to subsidize the process of discussion and preparation of the current *National Plan for Solid Waste* (*Plano Nacional de*

Resíduos Sólidos), it should be noted that the contributions of the gatherers in the area of recycling can and should be estimated economically and environmentally not inverting the hierarchy of values according to which "the social benefits, particularly the generation of employment and income for a portion of the population, a category of work before neglected, are the main factors that justify a public policy in favor of collectors of recyclable" (Freitas, Fonseca, 2012, p. 9).

While the recycling in a space of intersection of social and private interests, when does not affect by market structures inefficient as the oligopsonic, it is necessary to consider that, in addition to the opportunity costs, competitive and cumulative environmental advantages has caused a point turning not only national, but international, in this sense. It is becoming increasingly clear the importance of recycling for the minimization of the cost of production of a range of industries and the reduction of energy consumption, increasing its competitiveness, as well as its relevance to the retraction of the pollution of water, air and soil, preventing the flooding by blockage of manholes and trenches, thinning of landfills, sanitation of lakes, rivers and beaches, with impact on tourism, health and welfare social. Not really new, we already have more than twenty years of initiatives.⁸

The regulatory recommendation is indicative of the administration of minimum prices (or subsidized), whose financial costs are offset by marginal revenue derived from the differentiation of services able to retain customers on the basis of total utility of contrast agent local regulator, the example, in Salvador, the *Cooperative Complex Recycling of Bahia* (CCRB). Thus, the services there were differentiated by utilities logistical, operational and support to aggregate demand of the CCRB, which are the equal pay, exact and at the time of delivery, all gatherers, complainants balance vitiated in market controlled by agents intermediaries; the new infrastructure (organization, digital weighing, water, lighting, security, presence of SETRE/SESOL⁹ etc.); the personal protective equipment (EPI): uniform, boots, sleeve, bone, and ear protector; food three times a day and mineral water; collection points at strategic locations (Bahia, 2010); Credisol, new form of "credit solidarity" for the purpose of enlarging the capital in cash and the dilation of the ability to hold up to the cyclical revaluation of solid waste awaited for the post-Carnival (Bahia, 2010). The focus is on deletion of unequal exchange. The intermediaries of oligopsony could only compete with strategy of price of short-term, that the CCRB can respond in one of the movements, raising the price of one kilogram of aluminum can of R\$1.00 to R\$1.30, almost increasing tenfold billing physical, and, in the majority of the movements, raising of R\$1.30 to R\$1.50 the minimum price, with results also expressive, movements that are equivalent, in Graph 1, the variations

⁸ "The first programs of selective collection and recycling of solid waste in Brazil began in the mid-1980s to 1980, as innovative alternatives to reduce the generation of solid household waste and stimulation of recycling. Since then, communities organized, industries, companies and local governments have been mobilized and induced the separation and classification of waste in its sources. Such initiatives represented a great advance in respect of solid waste and its production. The first official information on the selective collection of solid waste were raised by PNSB 1989, which identified, on that occasion, the existence of 58 selective collection programs in the Country. This number grew to 451, according to PNSB 2000, and to 994, according to PNSB 2008" (Brazil, 2008).

⁹ The CCRB is composed of five cooperatives, all with a number of members between 20 and 26. The community involvement, however, goes far beyond them, extending over the 11 years of operation, the 18 (eighteen) thousand scavengers and partner associations bidding of food and clothes. Only in 2014, 15 groups of women co-joined the action, among other partners such as the Center of Art and Environment (Cama).

of p_1 to $p_2 - np$ and q_1 to $q_2 - qm$, where np and qm are the residual differential between the prices and quantities effectively practiced and that would eliminate the power monopsonic, i.e., when these m and $n = 0$, the goal higher.

Table 1: Social Indicators

		INDICATORS	Brazil	South	Southeast	Northeast	C. West	North	
CATEGORIES		Average age of scavengers	39.40	38.90	40.60	38.30	40.00	36.50	
		Percentage of women	31.10	34.10	30.90	29.30	34.10	29.10	
		Percentage of blacks (blacks and mestizos)	66.10	41.60	63.00	78.50	71.30	82.00	
	1. Demographics	Percentage of scavengers living in urban areas	93.30	93.50	96.20	88.50	95.60	93.20	
		Child dependency ratio in households with at least 1 collector	50.00	53.50	43.60	55.30	46.30	64.10	
		Average income from work of scavengers	571.56	596.90	629.89	459.34	619.00	607.25	
			Income inequality among pickers (Gini Index)	0.42	0.42	0.39	0.43	0.37	0.42
	2. Employment and Income	Percentage of residents in households with at least 1 collector extremely poor (less than \$ 70.00 per capita or US\$ 30,43)	4.50	4.10	2.20	8.40	1.80	3.80	
		Percentage of scavengers with social security contribution	57.80	59.10	56.10	61.50	55.10	54.80	
	3. Pension Plan	Coverage of the elderly population in entering with at least 1 collector	57.80	59.10	56.10	61.50	55.10	54.80	
		Illiteracy rate among pickers	20.50	15.50	13.40	34.00	17.60	17.20	
	4. Education	Percentage of scavengers with 25 years or more with at least elementary school complete	24.60	20.60	28.30	20.40	23.90	30.00	
		Percentage of scavengers with 25 years or more with at least secondary education complete	11.40	7.90	13.50	9.70	10.80	14.00	
		Percentage of children (0 to 3 years) living in households with at least 1 collector who attends daycare	22.70	19.80	27.90	21.70	18.50	13.00	
	5. Access to Public Services	Percentage of household with at least 1 collector with access to electricity	99.00	98.50	99.70	98.40	99.50	98.40	
		Percentage of household with at least 1 collector with a adequate sanitation	49.80	40.90	75.40	32.50	28.00	12.30	
	6. Digital Inclusion	Percentage of household with at least 1 collector with access to computer	17.70	20.10	26.40	7.00	19.20	9.00	

There is an existing diagnosis of Ipea, on the basis of the most recent Census (IBGE), concerning the current social situation of recycler (Silva, Goes, Alvarez, 2012, p. 8), which Table 1. The *expected improvement*, conservative, is tripling of average national income, duplication of pension contributions and reduction of 3/5 of the standard-deviation regional these two indices. With the integration of associations partners in supply of uniforms etc., it is expected a multiplier effect of 1 to 10 poor in metropolitan communities (*Botton Up*). The technology with chip sets streamline and cheapens the services collection and recycling. The economies with the prevention of floods and urban pollution in general are presumed. The revaluation of the professional recycler is the main cultural change expected, change slowly, and not without support of campaigns re-educative of population. *Needs and skills* of the project meet internally, by the actors involved. Each agent-local regulator must go through a process of incubation for 12 to 24 months, involving management training, financial, market strategies and cooperative principles.

The core of this project is the *reverse logistics of solid waste* (in the direction it has for the purposes of new Federal Law n° 12,305, Art. 11°: an instrument of *economic and social development characterized by the set of actions, procedures and means to facilitate the collection and refund of the solid waste to the corporate sector, for reuse in your cycle or other productive cycles* (or other final disposal environmentally appropriate, without prejudice to the systems of public collection of waste, *latu sensu*). This legitimacy did not exist. It gives legal support to initiatives to promote the selective collection. The garbage dumps were with your days counted since the institution of the

National Policy for *Solid Waste* (PNSR)¹⁰ in 2010, when states and municipalities had set 2012 as legal deadline

[...] to submit a plan for the management of solid waste, and only then, receive funds from the Union to work in this area. There will be no obligations to consumers, retailers and manufacturers. All will be subject to penalties of Law of Environmental Crimes if not reverse correctly the products after the consumption. The factories, for example, will have to collect the "excess residual waste" after use. The manufacturers of products with greater environmental degradation (pesticides, batteries, fluorescent lamps, tires and electric-electronic) are obliged to implement systems that allow the recollection of the products after the use by consumers. The text creates the so-called "reverse logistics" (Article 3º, item XII) for collection of discarded products by consumers. Dealers and distributors will be the main points of fencing of discarded products, which should then be sent to the manufacturers or importers. The latter will give the final destination to the trash. (Denadai, 2010).

Experts of the Ipea concluded that, in the actual pace of operations, it will be very difficult achieve the goal of extirpating the garbage dumps until 2014, non-compliance regardless of which, once this deadline has passed, garbage dumps will be considered illegal and, for any purpose, "outside the law". The law establishing the *National Policy of Solid Waste*, in its Art. 11º, establishes that the selective collection system should prioritize the participation of cooperatives or other forms of association of collectors of reusable materials and recyclable consisting of persons of low income. The connection between this sector and the public policies directed at associations is new.

The first paragraph of Article 18º of lays down that the participation of the gatherers in systems of reverse logistics should be prioritized, particularly in the case of post-consumer packaging, as well as, according to the Art. 23º, the sectorial agreements aiming at the implementation of the reverse logistics should contain the possibility of hiring organizations of gatherers for the implementation of the proposed actions on the systems to be deployed. States and municipalities legislate, in a complementary way, in order to apply the proper law of the Solidarity Economy, as in this nine cases of Pernambuco, Rio G. do Norte, Santa Catarina, Minas Gerais, Mato Grosso, Rondonia, Mato G. do Sul, Acre, Bahia, four other projects-of-law state and numerous legislative initiatives municipals. In June 21, 2006 was published the *Decree nº 5.811* that had on the composition, structure, powers and functioning of the National Council of Solidarity Economy - CNES. Finally, tying legally the three variables in this project (recycling, cooperatives and fair-price), the signing, by the *Presidency of the Republic*, the *Decree nº 7.358*, of November 17, 2010, instituted the *National System of Fair Trade*¹¹, becoming the first country with a such system, which redefined:

¹⁰ Sanctioned by the *Presidency of the Republic* after being approved in a joint meeting of the CCJ (*Commission Constitution, Justice and Citizenship*), CAE (*Commission of Economic Affairs*), CAS (*Commission for Social Affairs*) and CMA (*Commission of Environment*) of the Senate, this Federal Law no. 12.305/2010 establishing the *National Policy for Solid Waste* (PNRS), regulated by Decree nº 7.404 (December, 23/2010), it was processed in *National Congress* for 21 years.

¹¹ "The National System of Fair Trade (SNCJS) is a set of parameters: concepts, principles, criteria, actors, instances of control and management, organized in a single strategy of affirmation and promotion of Fair Trade in our country. Organized in a document that merges regulatory mechanisms and incentives, the SNCJS seeks to consolidate as a public policy, through the enactment of a law that would institutionalize. It is, thus, as political project as economic. Political, because it formalizes the recognition by the Brazilian State of Fair Trade as social policy of confronting social inequalities and the precariousness of labor relations. And, economic, to provide an identity to the products and services of the Solidarity Econ-

I - fair trade: commercial practice differentiated based on the values of social justice and solidarity held by ventures economic solidarity; [...]; V - fair price: is the definition of value of the product or service, built from the dialog, transparency and effective participation of all the actors involved in its composition that results in equitable distribution of gain in the productive chain. *Unique Paragraph* - The terms fair trade, fair exchange, fair trade, fair equal, alternative trade, solidary trade, ethical trade, ethical trade and solidary are understood in the concept of fair trade, under the terms of this Decree. (Brazil, 2010, italics originals).

4. BACKGROUND, GOALS AND STRATEGIES OF THE POLICE

The policy proposed here is directed to Brazilian Metropolitan has been verified the exercise of power oligopsonicin reverse logistic of solid waste, having aim to establish new standard of competitiveness and fairness intra and inter-regional, combining innovative efficiency allocative and distributive to technologic innovation local of the sector by *System of Radio Frequency Identification (RFID)*.

Although have been identified good tests in regulatory enforcement in the Country, as example of the Campaigns *The Decent Work Preserves the Environment*¹², developed in the state of Bahia¹³, the incipience of linkage intra-regional unfolds in a uncoordinated national policy in relation to systems of management of regional developments in sectors recyclers. For that year, which is a transverse action that alliance public authorities and civil society around commitments socio-environmental¹⁴, it should be stressed that

omy, adding value and concept to them, and, thus, expanding its sales opportunities." (Faces of Brazil, 2013). "Art. 3º -The SCJS is designed to strengthen and promote the concept of fair and equitable trade in Brazil, which includes achieving the following goals: I - strengthen national identity of fair trade, and through the dissemination of its concept, its principles and criteria for the recognition of practices of fair trade and its promotion; II - encourage the practice of fair price for anyone who produces, sells and consumes; III - disseminate the products, processes, services, as well as the experiences and organizations that must comply with the standards of the SCJS; IV - subsidize the economic solidarity ventures, accreditation bodies of conformity of this entities with the fair trade practices, assessment and the supporting entities promoters of fair trade, by means of the national basis of information on solidarity economy and economic solidarity ventures with practices of fair trade recognized by SCJS; V - help with public and private efforts to promote actions to promote the improvement of conditions for the marketing of economic ventures solidarity; VI - encourage the economic collaboration between economic solidary enterprises [...]." (Brazil, 2010).

¹² "The strategy of intersectionality in the implementation of the policy of support for cooperatives and associations active with solid waste has proved to be consistent in RMS (Metropolitan Region of Salvador) during the periods analyzed, covering aspects age, environmental, health, organizational, logistical, technological, competitive, lending and ethnic-cultural." (Primo, 2011, p. 36). Restructured from the report of the Public Prosecutor's Office in 2007, the policy of recycling the SESOL/SETRE became a comprehensive and transverse action regional involving state, municipalities, businesses, enterprises and civil society, removing hundreds of tons of waste from the environment and benefiting tens of thousands of collectors. Although the policy is re-territories for well beyond the RMS, she comprehend the municipalities of Camaçari, Candeias, Dias d'Ávila, Itaparica, Lauro de Freitas, Madre de Deus, Mata de Sao Joao, Pojuca, Salvador, São Francisco do Conde, São Sebastião do Passé, Simões Filho and Vera Cruz, influencing more directly in Sergipe, Alagoas, Pernambuco and Piauí.

¹³ 43% of ventures of solidarity Brazilian are in the North East, but only 10% are in Bahia (approximately 2.200 initiatives, involving around 80 thousand people). The other 90%, and of these 10%, interest us as potential beneficiaries.

¹⁴ The multiple needs of strategic design, marked the previous note, correspond coordinated efforts of the Secretariats of State Setre, Seats, Sedir, Serin, in addition to the Car, Casa Civil, Uneb, Ingá, UFRB, Abralatas, Petrobras, Brahma, Center of Art and Environment, Municipal Departments of Public Services (Sesp), Cleaning Firm Urban (Limpurb), in addition to the Hatchery for Economic Ventures and Supportive of the University of the State of Bahia (Incubation/Uneb), among other partners and several coopera-

at least since 2012, the *Brazilian Association of Producers of Highly Recyclable Cans (Abralatas)* "want to submit this project to other states, as an example of initiatives that are able to minimize problems social, environmental and economic in a single action", as note disclosed in February 23, 2012 by Ascom/Setre, but without corresponding result multiplier for the other regions of the country, so that it is still, in general, unique. On 14, 2012, likewise the *Central Cooperative Ventures and Solidarity of Brazil (Unisol - Brazil)* did record that "is engaged in the project to strengthen it further and, in particular, contribute to the socialization of experience", in expectation. Assuming that general objective, the policy proposed here has the following specific objectives:

- Eliminate nationally the *Exploitation Oligopsonic* of recyclers (Trajectory of the Minimum price to *Fair Trade*) integrating regional markets through the federal administration of the price of each element recyclable, safeguarding standards in competitive reverse logistics of solid waste;
- Raise significant and persistently the amount of recycled solid waste industrially useful and raise the level of *Environmental Sanitation* in the short term;
- Educate the public through incentives (prizes) and burden (system of fees and surcharges) through monitoring pilot their discards of solid waste through the chip RFID (radio frequency identification).

The first step of the project is to characterize, so calculated, the market structures of solid waste in metropolitan areas of all regions, by estimating the corresponding deviations between prices and values added, privileging cyclical and structurally the links in the chain of reverse logistics of the residue from the scavenger loose to industry. It is justified, because the fight multilateral of civil society and the public authorities against the distributive inefficiency-allocative resulting from the merger is theoretically based between us:

The trend toward concentration that created in certain branches situations of virtual monopoly provoked reactions reversed in defense of the public interest, such as the anti-trust laws the end of the last century. Closed the door to the monopoly, it was necessary to develop ways to coordinate more subtle. The oligopoly [and oligopsony] is the culmination of this evolution: it allows a small group of large firms to create barriers to entry for other in a sector of economic activity and administer jointly the prices of certain products, preserving, however, autonomy financial, technological and administrative. The administration of prices creates relative advantage for companies that more innovate both in productive processes and the introduction of new products in a given sector. Differently of traditional competition of prices which translates into reduction of profits, weakening financial, plant closures or, in the case of empires a monopolist, increase in prices and demand reduction, the world of the oligopolies resembles more the [...] a sport to which only has access champions. (Furtado, 1996, p. 26-27).

Already in *The Economy and the Art of Controversy*, K. Galbraith intervened in debate of *unequal exchange*, focusing on the situation of the small farmer¹⁵, but it is a dialectic

tives in the state. The *transversally, territoriality and the thematically* constitute the so-called "3T" of Plan Pluri-Anual – PPA at 2012-2015.

¹⁵ "The struggle in the labor market, with its controversy, was the result of primitive inequality in ability to contract and the efforts of unions to delete it. This inequality is not exclusive to the labor market, nor are the consequences to which it gives rise. Similar Problem occurs there where many people, individually powerless because they are numerous, exchange with some large companies, of relative strength much higher, precisely because they are few. This is the weakness of the lone worker, who also is characterized, in general, the situation of the farmer. So typical, the farmer alone does not influence on the price of sale

relapsing in Political Economy. In his free-teaching thesis submitted to the Institute of Economics at the State University of Campinas (Unicamp), M. Coutinho has recovered the classics, more particularly of Adam Smith, which, in particular without-had, "the question of the value should refer to the faculty that the goods have, in exchange, to command the labor of others." (Coutinho, 1993, p. 113). But, regardless of whether it is measured by work or by another measure, the issue is that the value may differ from the price and, in this case, that profits are transferred winner of the loser. In the most recent global stage of mergers/acquisitions, the policy measures for the prevention of the abuse of power in systems of trade, though not restricted to oligopolies, whose action is free, have become commonly used, including in Brazil, and it is opportune to mention here in note¹⁶. Also called the "degree of monopoly", the *market power of the company* can be kaleckian defined, so simple that was, once, even considered tautological, where $K = \frac{p}{u}$ in which p is the *price*, u direct costs and K the *degree of monopoly*, which changes due to 1) changes in the degree of economic concentration, 2) advertising (and middlemen), 3) The ratio of indirect costs (wages, cost of capital, taxation etc.) on the direct and 4) power of the trade unions/cooperatives. Recalling that the direct costs (u) are data for the costs with wages (w) and by expenditures (both measured per unit of product) with raw materials (m), so that $u = w(1 + j)$, where j is the quotient raw materials by the hourly wage, come

$$K = \frac{p}{w \left(1 + \frac{m}{w}\right)}$$

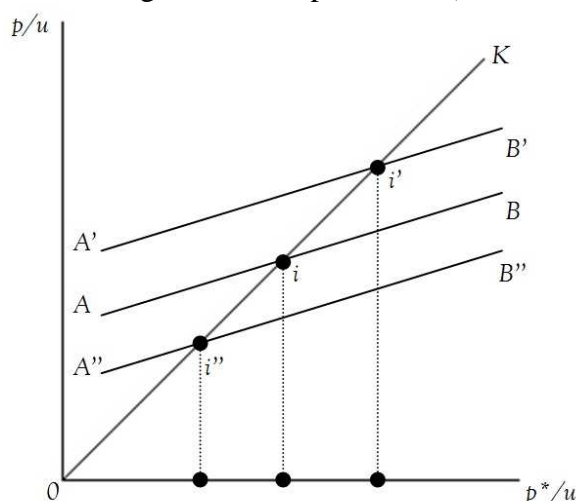
It stands out in this brief argument to influence the degree of monopoly on the formation of prices, each firm in an industry more or less concentrated by fixing prices by mark-up.¹⁷ Low prices, with subsequent restriction of purchases, is inversely propor-

or purchase. He lives in a world where, as a result of stronger position occupied by other circles on the market, it has become commonplace the exercise of some power to influence prices. [...] Approaching its weakness in the ability of the contract, the worker never had doubts about the appropriate solution, which was to organize a union. The vision of the farmer of his problem and its solution has never been so defined. [...] Voluntary Organizations to expand its economic power - cooperatives - seemed at moments constitute the solution." (Galbraith, 1959, p. 36-39). For a theoretical comprehension more complete the *task* of development, see Furtado (1979).

¹⁶ "After the democratization process, more specifically from the year 1994, the date of the promulgation of the new antitrust law Brazilian, the Brazilian System of Defense of Competition - SBDC began to become apparent through analysis of acts of concentration. Many famous trials was press in memory, as paradigmatic examples, as the acts of concentration involving companies Kolynos and Colgate (AC no. 0027/1995), Antarctica and Brahma - AMBEV (AC no. 08012.005846 / 99-12) and, more recently, Nestlé and Garoto (AC no. 08012.001697 / 2002-89). To analyze acts and contracts that may have an impact on competition, in accordance with the procedure laid down in Article 54º, caput and paragraph 3º of Law no. 8.884, components of SBDC exert its preventive function, whose objective is to prevent the possible formation of market power that will enable a possible abuse more ahead by the company resulting from. [...] The components of the SBDC also have legal competence to repress anti-competitive behavior, which constitutes its repressive function. In fact, after a start widely focused on analysis of acts of concentration, from 2003 it started a strong trend toward focusing on the actions of public officials in the fight against cartels, especially with the effective use of agreements of leniency, as well as advanced techniques of research, such as the search and seizure and the telephone interception. In order to punish the anticompetitive conduct proven, components of SBDC may impose financial penalties and not cash too." (Ragazzo, 2007, p. 1-2).

¹⁷ A normal feature of the economy, second Kalecki, is operating with idle capacity due to the degree of monopolization of the economy. Thus, the production is not where the price equals the marginal cost, but which gives the maximum profit. With the existence of idle capacity and given the characteristics of the production, there is a follow-up on horizontal curve marginal cost, and the price determined by a margin on costs according to a rule of *mark-up*. (Gremaud, Braga, Vasconcelos, Toneto Jr., 2000, p. 142).

tional to the competition. This rule is well-known, so that the price at which the companies offer their goods and services is a function of their direct costs and the price of its competitors, in the form $p = mu + np^*$ in which p is the price, p^* is the price of the good or service substitute (weighted average of prices of all firms competing within the structure of the market), u are direct costs and m a positive margin (*mark-up*) on these costs, by order, n being a coefficient of market power within the range $0 < n < 1$, so that the smaller the size and the market power of the Company, p^* both more converges to p and the quantities produced more approaching the level that could be achieved were it not for the influence of the degree of monopolization (Kalecki, 1976).



Graph 2: *Modification of the Degree of Monopoly*
Source: IE-UFRJ, Fonseca (2014).

By dividing the formula of rule *mark-up* by u we obtain the following equation of a straight line in which m converts the linear coefficient of an affine function, vertical effect indicated in the graphic above:

$$\frac{p}{u} = m + \frac{np^*}{u}$$

Plotting in an axis coordinated p/u and p^*/u it is possible to observe (Graph 2) that maintained constant direct costs (u), when m grows, that is, when the degree of monopoly rises, the straight AB moves upward ($A'B'$), or, if decreases, down to ($A''B''$). This representation adopt the line of 45° (OK) as the axis on which $p/u = p^*/u$ and by which are marked i , i' and i'' whose project on the abscissa is directly proportional to the degree of monopolisation. In the stretches of straight lines AB to the left of the shaft OK the ratio of prices to direct costs of the firm is greater than the set of competitors. It is not here exhausted the possible inferences of this model, when vary u emprise to emprise or its market power, or when m is equal to the average degree of industry, converting the firm in industry representative. Knowing that the smaller the number of firms and the more unequal distribution of size between them, less competitive and more concentrated is the market, there are many ways to measure m and n , among which is the method *Herfindahl-Hirschman Index* (HHI) used in the United

States in one of its variants, of which the usual formula of *Hannah and Kay* (1977) is the general case, index mathematically given by

$$HK_{NE}(\alpha) = \left(\sum_{i=1}^n S_i^\alpha \right)^{\frac{1}{1-\alpha}}$$

where S_i is the *market share* (the percentage of the market/industry of each i -th firms) n is the number of firms, $0 < \alpha \neq 1$ and the subscribed *NE* indicates that its result is calculated in such a way as to present the response in terms of the *Equivalent Number*. An orientation to the use of Excell for the calculation of this index for any industry can be found in P. Latreille and J. Mackley (2014). What matters here highlight is just that m and n are determinable and are often known, and that can and should fall, where are taken into account the interests of the company, offended by shortage speculative and/or artificially more expensive goods and services, since any *market power* m or n . On the basis of these calculations it is possible to identify possible environment to abuses on the part of the industry demand for solid waste in Brazil.

A second step is the *innovative proposal* to connect directly the networks of cooperatives to the industries, through contracts, with the stipulation of minimum selling prices (on the part of recyclers associated with) for each residue, adjusted quarterly by *tripartite commission and joint representatives of public authorities, civil society and business*. The goal is to define values of products or services - from the dialog, transparency and effective participation of all the actors involved in its composition - generating the trade resulting in equitable distribution of gain in productive chain, in conformity with Art. 2° of Decree nº 7.358. This law is a new and advanced instrument to solidarity.

5. LEAP IN TECHNOLOGY

Technologically, it is a situation of almost zero incorporation of machinery, equipment and technology of information to the processes of selective collection and recycling, in the vast majority of cases surviving situations of garbage dumps in outdoor and spurious trade. The fourth and last variable of present project is a *technological innovation* capable of leveraging the selective collection through incentives and over-taxation electronically monitored, accelerating and expanding the system, possibility open actually.

This second front is constituted by the deployment-pilot, in metropolitan regions of two systems of chip sets of gatherer boxes junk RFID (radio frequency identification) 1) individual and 2) collective, with the control of who discards what and how much. This technology has been considered a success in the Dutch city of Enschede, where recycling has grown 42%, making the educational campaigns more effective. It can run a dual vector of stimuli of good practices and inhibition of bad-conduct in relation to the production of irrational solid waste, because, relating the garbage to those who produce it can, according to the *U.S. Environmental Protection Agency*, offer bonus to consumers aware, for purchases in shops accredited, as done in Chicago or, conversely, encumber with a surcharge that raising the services of waste collection, as done in Maia, Portugal, informs the *Group of Studies of Spatial Planning Environmental Portugal*. The technology deserves special attention in edition of dec./2010

of *Superinteressante* review, which synthesized the efficiency and low-cost of RFID systems, which uses the strategy of containers underground and personal card. Of reason cost-benefit low, the system will refine and restructure the studies measured gravimetrically and nationally in sanctions and rewards, and will generate a gradient of convergence between the standards of domestic generation of municipal solid waste, state and regional, directly proportional to its progressiveness in relation to the income in each situation of bad conduct of selective discard.

The cost-effectiveness of the system is low because reduce to less than half of the current expenditure logistics of collection. The system is flexible, account with numerous technical options and is adaptable to local needs. For example, it can treat differently the more thinly populated regions and central, as already done in the aforementioned city Enschede.

Each property (in the streets with relatively few households) receives a set of cans of different colors - one for each type of waste. They are at home, and they are only made on the street on the dates determined by the company's calendar collection. The brass is installed a chip RFID (radio frequency identification). He works as a barcode: saves a number of users are then crossed with a database. The difference is that it can be read up to 5 meters away. When the trash truck raises the brass, your RFID reader identifies the user and records the data collection: when and where it was made and what the weight of the waste. Then, it sends the information to the company. [...]. For central regions: The company pickup installs large containers of garbage in the basement of public areas and distributes to each home a card with RFID chip. The user can make the disposal of garbage on the day that he wants to. The cover of the container opens when the resident is approaching with the card. Data such as user, date of discharge and weight of garbage are stored on the card or transmitted to the company collection. As each container meets multiple users, the trucks do not need to stop all the time for collect the garbage. Door-to-door turns on a block-by-block basis. This leaves the collection much faster. (Steffen, Cunha, Sorano, Ducroquet, 2010).

6. DANGER, PARTNERSHIPS AND SUSTAINABILITY

From the social point of view, the main factor is the possible existence of indices of income inequality within the cooperative, which requires attention to the distribution of rents. The articulation and coordination of the efforts of public authorities and private are the central challenge, but the new regulatory frameworks requires the adoption of measures protective that mitigate risk. First of all, the Ministry of Development and Regional Integration need to follow narrowing their policies with those of the Ministry of Labor, through the National Secretariat of Economy of Solidarity, and both have to strengthen its ties with the State Secretariats of Labor and the National Forum of the Secretaries of Labor - FONSET, until the corresponding Municipal Secretaries of Labor. Giving sustainability executive this first institutional arrangement, follow with the Casa Civil, the Ministries of the Cities and Environment, Science and Technology, Secretariats of Development and Social Assistance, Child Protection, Justice and Citizenship, by supposed, always in conjunction with other national bodies and state related to the activity of recycling, such as the Fund for Combating and Eradicating of Poverty - FUNCEP, Development Agencies and Public Banks able to offer encouragement and claims solidarity (differentiated for cooperatives), finally, extending the current partnerships with NGOs and companies of wide acting as Abралatas, Cempre, Petrobrás, Brahma, Coca-Cola, Unisol, Pangaea and Caritas, as well as with the Forums and Councils National, State and Municipal of Solidary Economy etc. The goal is to make the growth and convergence of rent.

The project is backed by three relatively new Federal laws, by nine recent state laws and numerous municipal laws that proliferate, being strictly convergent to the plans of government and to the priorities of the 27 states of the Federation, offering stable and increasing opportunities for returns. The empowerment (self-management) of cooperatives, done well, will make the system turn without cost in 36 months, opening from there, a virtuous cycle new reverse logistics.

7. SIMILAR EXPERIMENTS, DEPLOYMENT TIME AND REPLANNING

With regard to the aspect Technologic-Operational of project, pioneering experiences and successful in Brazil with chip sets garbage containers and underground were carried out, in 2012, with prospects for wide dissemination in 2014, at the intersection of Brigadeiro Faria Lima avenue and Rebouças, in the city of São Paulo, separating organic Waste from recyclable, as well as in the district of Parada de Taipas (North Zone) and in the Municipal Market, in addition to other 27 locations then planned, whose collection ability, set out in the contract with the concessionaire responsible for testing, is 165 thousand tons of waste until 2019. Was in city of Caxias do Sul that took place one of the first experiences with this kind of technology in Latin America, through the Collection System Mechanized Side, deployed by Portuguese company *Sotkon Waste Systems Brazil*, in Street Alfredo Chaves, to demonstration; this experience, however, remained isolated.

As to the Social-Technologic aspect of the present project, regulatory experience pioneering and successful has been recorded over the 11 years of Recycling Campaigns in Bahia, through the SESOL/SETRE, Prefecture, NGOs and private sector, as we have already seen. As to the costs of manufacture and installation of equipment for *recycling system with chip*, it can be shared between the actors involved, as each scenario of relations between the public and private spheres, but the implementation needs to be owned by their territoriality. Must begin with the largest urban centers, in the first phase, scaling the other, being feasible duplicate the recycling in 18 months.

The project has economic returns and social immediate, with strong potential for expansion in the middle and long term, on the one hand, on the basis of pent-up demand for recycling of waste and, on the other hand, according to be passing through a time of technological and regulatory inflection; the maturity, in this first stage, it is expected for 36 months, with the first 12 to 24 months for the process of incubation and qualification (strategic) of the agent regulator local, on the one hand, and the creation and deployment of the system of recycling with chip sets, on the other hand. In addition to this period, more 12 months of training with new technologies, adjustments of trajectories or functions-reaction to regional actors-competitors and the cyclical fluctuations per annum, providing opportunities studies of measured gravimetrically and inspection *in loco* for effective verification of the results obtained, reevaluating scenarios, redesigning partnerships and raising the sustainability of the changes triggered as a whole.

REFERENCES

- Agência Brasil. (2009). *Menos de 2% dos resíduos sólidos produzidos no Brasil são reciclados*. Recuperado de <http://ultimosegundo.ig.com.br/ciencia/meioambiente/2013-09-09/menos-de-2-dos-residuos-solidos-produzidos-no-brasil-sao-reciclados.html>.
- Ascom-Setre.(2012). *Campanha 'Trabalho Decente Preserva o Meio Ambiente' coleta 65 toneladas no Carnaval de Salvador*. Recuperado de <http://homologia.setre.ba.gov.br/noticias/campanha-trabalho-decente-preserva-o-meio-ambiente-coleta-65-toneladas-no-carnaval-de-salvador-3>.
- Bahia. (2010). Secretaria do Trabalho, Emprego, Renda e Esporte. Superintendência de Economia Solidária (SETRE/SESOL).*Relatório Síntese das Políticas Públicas de Economia Solidária Desenvolvidas Pelo Governo do Estado da Bahia: Período 2007-2010*. 2010. Salvador: SETRE.
- Bettelheim, C. (1972). “Observaciones Teóricas”, en A. Emmanuel. *El Intercâmbio Desigual: Ensayo Sobre los Antagonismos en las Relaciones Internacionales*. México: Siglo XXI.
- Bosi, A. (junho/2008). “A Organização Capitalista do Trabalho ‘Informal’: O caso dos catadores de recicláveis”, en *Revista Brasileira de Ciências Sociais*, v. 23, n. 67.
- Brasil.(2010). Presidência da República.Casa Civil.*Decreto Nº 7.358, de 17 de novembro de 2010. Institui o Sistema Nacional do Comércio Justo e Solidário - SCJS, cria sua Comissão Gestora Nacional, e dá outras providências*.
- Brasil. (2008). Ministério do Planejamento, Orçamento e Gestão. Instituto Brasileiro de Geografia e Estatística – IBGE.Diretoria de Pesquisas Coordenação de População e Indicadores Sociais.*Pesquisa Nacional de Saneamento Básico 2008*.Brasília: IBGE.
- Carmo, M. (2011).“As Políticas de Estimulação ao Trabalho dos Catadores no Rio de Janeiro Como Estratégia Oriunda da Valorização dos Resíduos”, en *Cadernos Gestão Pública e Cidadania*, v. 16, n. 58. São Paulo.
- Carmo, M. (abr./jun. 2009). *A Semântica ‘Negativa’ do Lixo Como Aspecto ‘Positivo’: Um Estudo de Caso Sobre uma Associação de Recicladores na Cidade do Rio de Janeiro, Brasil*. Administração Pública e Gestão Social, Viçosa, v. 1. n. 2, pp. 21-50.
- Chiang, A. (1982). *Matemática para Economistas*. São Paulo: Ed. USP.
- Coutinho, M. (1993).*Lições de Economia Política Clássica*. São Paulo: Editora Hucitec.
- Dagnino, R y R. Dagnino. (julho/2010). “Políticas para Inclusão Social de Catadores de Materiais Recicláveis”, en *Revista Pegada – especial*, n. 66.
- Denadai, N. (2010). *Presidente Lula sanciona lei que prevê fim dos lixões*. Recuperado de <http://www.projetomundomelhor.com/2010/08/s-era-o-fim-dos-lixoes.html>.
- Faces do Brasil. (2013). Plataforma de Articulação do Comércio Justo e Solidário. *Conheça o SNCJS*.Recuperado de <http://www.facesdobrasil.org.br/sistema-nacional-de-comercio-justo-e-solidario/3-sncjs.html>.
- Fernandez, J. (2012): *Diagnóstico dos Resíduos Sólidos de Logística Reversa Obrigatória*: Relatório de Pesquisa. Brasília: Ipea.
- Freitas, L. y I. Fonseca. (2012). *Diagnóstico sobre Catadores de Resíduos Sólidos: Relatório de Pesquisa*. Brasília: IPEA.
- Furtado, C. (1996). *O Mito do Desenvolvimento Econômico*. Rio de Janeiro: Paz e Terra.
- Furtado, C. (1979). *Teoria e Política do Desenvolvimento Econômico*. São Paulo: CEN.
- Galbraith, K. (1959). *A Economia e a Arte da Controvérsia*. Rio de Janeiro: Fundo de Cultura.

- Gremaud, A. y M. Braga y M. Vasconcelos y R. Toneto Jr. A Economia Kaleckiana e o Princípio da Demanda Efetiva. In. Lopes, L.; Vasconcelos, M. (Orgs.). *Manual de Macroeconomia: Nível Básico e Nível Intermediário*. São Paulo: Atlas, 2000.
- Grimberg, E. y L. Tuszal y Y. Goldfarb. (2004). *Gestão Sustentável de Resíduos Sólidos e Inclusão Social: Estudo de caso sobre as Cooperativas das Centrais de Triagem do Programa Coleta Seletiva Solidária da Cidade de São Paulo*. São Paulo: ITCP/USP.
- Harris, D. (1974). “Um ‘post mortem’ à Parábola Neoclássica”, *Pesquisa e Planejamento Econômico*, Rio de Janeiro 4(3), pp. 563-600, dez.
- Kalecki, M. (1976). *Teoria da Dinâmica Econômica*. São Paulo: Abril.
- Lange, O. (1986). Algumas Observações Sobre a Análise Insumo-Produto. En. Lange, O. *Ensaaios Sobre Planificação Econômica*. Nova Cultural: São Paulo.
- Leftwich, R. (1974). *O Sistema de Preços e a Alocação de Recursos*. São Paulo: Pioneira.
- Latreille, P. y J. Mackley. “Using Excel to Illustrate Hannah and Kay’s Concentration Axioms”, *International Review of Economics Education*. Recuperado de <http://www.heacademy.ac.uk/resources/detail/subjects/economics/10-1-Using-Excel-to-Illustrate>.
- Lopes, T.; Neder, H. (2011). Teoria da Programação e Economia Política: uma introdução alternativa a partir de Sraffa, Leontief e Lange, *Anais*, 39º Encontro Nacional de Economia/ANPEC, 2011. Recuperado de <http://anpec.org.br/encontro/2011/inscricao/arquivos/000-b63a24aa275f5c5156d906357513a62f.pdf>.
- Magera, M. (2008). *A Reciclagem dos Resíduos Sólidos Urbanos e o Uso das Cooperativas de Reciclagem: Uma Alternativa aos Problemas do Meio Ambiente*. Recuperado de http://www.fundace.org.br/cooperativismo/arquivos_pesquisa_ica_la_2008/043-magera.pdf.
- Martins, L. y M. Peres. (24 to 26 of october, 2006). “Economia Solidária: A Experiência da Cooperativa do Reciclador Solidário de Piracicaba”, en Unimep. *Anais*. 4ª Mostra Acadêmica.
- Marx, K. (1996a). *O Capital: Crítica da Economia Política (O Processo de Produção do Capital)*. Livro Primeiro, Tomo 1. São Paulo: Editora Nova Cultural.
- Marx, K. (1996b). *O Capital: Crítica da Economia Política (O Processo de Produção do Capital)*. Livro Primeiro, Tomo 2. São Paulo: Editora Nova Cultural.
- Marx, K. (1991). *O Capital: Crítica da Economia Política (O Processo de Circulação do Capital)*. Livro Segundo. São Paulo: Bertrand Brasil.
- Marx, K. (1987). *Teorias da Mais Valia: História Crítica do Pensamento Econômico (Livro 4 de O Capital)*. Volume 1. São Paulo: Bertrand Brasil.
- Pasinetti, L. (1969). “Switches of Techniques and the ‘Rate of Return’ in Capital Theory”, *Economic Journal*, vol. 79, pp. 508-31.
- Primo, R. (2011). “Política Transversal de Regulação do Mercado Oligopsônico de Resíduos Sólidos nos Carnavais de Salvador de 2007 a 2012 por meio de um Complexo de Economia Solidária”, *Isegoria - Ação Coletiva em Revista*, UFV - Área Temática: Economia Solidária e Cooperativismo, Ano 1, vol. 1, n.2, set. de 2011/fev.
- Ragazzo, C. (2007). *Análise Econômica da Regulação: O Papel da Advocacia da Concorrência. Berkeley Program in Law and Economics*. UC Berkeley: Latin American and Caribbean Law and Economics Association (ALACDE) Annual Papers.
- Ricardo, D. (1988). *Princípios de Economia Política e Tributação*. 3ª ed. São Paulo: Nova Cultural.
- Robinson, J. (1953). “The Production Function and the Theory of Capital”, *Review of Economic Studies*, vol. XXI (2), nº 55, pp. 81-106.

Sachs, I. (2009). “A Saída da Crise: Novas Tarefas para RTS –Tecnologia Social, Sustentabilidade e Cidadania”, en *Rede de Tecnologia Social*.Tecnologias Sociais Caminhos para a sustentabilidade.Brasília.

Salvatore, D. (1984): **Microeconomia**. Trad. and rew. Marco Antônio Sandoval de Vasconcelos. São Paulo: McGraw-Hill do Brasil.

Samuelson, P. “Parable and Realism in Capital Theory: the surrogate production function”,*Review of Economic Studies*, vol. 39, pp. 193-206, 1962.

Sevcenko, N. (1999). *Literatura como Missão: Tensões Sociais e Criação Cultural na Primeira República*. São Paulo: Brasiliense.

Silva, S. y F. Goes y A. Alvarez (2012). Situação Social das Catadoras e dos Catadores de Material Reciclável e Reutilizável: Região Nordeste. Brasília: Ipea.

Steffen, R., Cunha, S., Sorano, V., Ducroquet, S. (dez. 2010):Como Funciona a coleta seletiva com chip?*Superinteressante*.

Smith, R. (1985). *Troca Desigual e Industrialização no Nordeste*. São Paulo: IPE.

Vilanova Neta, M. (2011).“Manejo de Resíduos Sólidos”, en IBGE.*Atlas de Saneamento 2011*. Brasília: IBGE.

APPENDIX A - *Unequal Exchange and Equilibrium Asymmetrical*

A.1 Diagrams Adapted Marx's Writings: Preliminary Considerations

Here we describe the economic model with unequal trade most concise possible, covering a minimum number of sectors and disregarding the existence of the government and foreign market. In section A.2 we will consider the system with n sectors.

Each residue enters in two forms:commodityto the consumption and/or to the production. Different the analysis of intersect oral reproduction of capitalist economic system made in *Book II* of *The Capital*(1991), but keeping with her a theoretical and mathematics identity, here it is considered that the value of the national product, during a given period (a fiscal year, for example), is composed of four parts:

- Two on the income of the owners of the workforce (employees direct, autonomous o cooperatives), already deducted the plot on the unequal exchange with the owners of the means of production, which are v and s^I , respectively, *salaries* and *surplus-value I* (work non-paid);
- Other two constitutive of the value of the means of production used (raw materials, inputs, depreciation of machinery, technological depreciation etc.) of ownership of suppliers intermediaries, already deducted from or added to the portion of the unequal exchange with the owners of the means of production consumers of their products (banks, industries, agribusiness, commerce, etc.), which are c and s^{II} , respectively, *production costs paid* and *surplus-value II* (production costs non-paid).

A primary condition for the sustainability of the system is that s^{II} does not compromise the refitting of capital c and, in fact, it is no longer that plot of s^I . The *cost of producing non-paid* are merely *transfers* of intersectoral *surplus-value I*. The *work non-paid total* (s) can be defined or subdivided, thus, as $s^I + s^{II}$. The *national product total* (P^N) can be represented by the sum of four plots:

$$P^N = c + v + s^I + s^{II}$$

Unifying gains of *unequal trade*, as *equilibriums asymmetrical* e^i , is accepted:

$$s^I + s^{II} = e^i$$

Segmenting the economic system integer in two major sectors, compounds by *producers of goods of production* or *intermediaries consumer* (P^N_1) and *goods of consumer or of final consumer* (P^N_2), it can be structured:

$$\begin{aligned} \text{Total output of goods of production} &= P^N_1 = c_1 + v_1 + e_1^i \\ \text{Total output of goods of consumer} &= P^N_2 = c_2 + v_2 + e_2^i \\ \text{National Product Total} &= P^N = c + v + e^i \end{aligned}$$

A.1.1 The Simple Reproduction

The dynamics of an economy without growth, that reproduces itself continuously, requires that the total demand of the means of production is equal to the total *needs of refitting* of the system, in the same way that the total demand for consumer goods is identical to the wage's fund and earnings totals with the *unequal trade national*.

$$\begin{aligned} \text{Total Demand of goods of production} &= c_1 + c_2 \\ \text{Total Demand of goods of consumer} &= v_1 + v_2 + e_1^i + e_2^i \end{aligned}$$

There is the *condition of equilibrium 1*, *identity of demand and product* of the means of production:

$$c_1 + c_2 = c_1 + v_1 + e_1^i \quad (1)$$

There is the *condition of equilibrium 2*, *identity of demand and product of means of consumption*:

$$v_1 + v_2 + e_1^i + e_2^i = c_2 + v_2 + e_2^i \quad (2)$$

Simplifying the equations 1 and 2, we have, respectively, these identical equations 3 and 4:

$$c_2 = v_1 + e_1^i \quad (3)$$

and

$$v_1 + e_1^i = c_2 \quad (4)$$

Are identical thus indicate the *relation input-the product* of this *double-sectoreconomy*, such that:

$$\begin{aligned} P^N_1 &= c_1 + v_1 + e_1^i \\ P^N_2 &= c_2 + v_2 + e_2^i \end{aligned} \quad (5)$$

Analytically, the product of the Sector 1, whose value is equal to c_1 , remains within the sector, to *refitted to the means of production consumed*; the remainder, **in bold**, with a value equal to the sum $v_1 + e_1^i$, is replaced with c_2 from Sector 2 that, reciprocally,

preserves $v_2 + e_2^i$ within itself, to its final consumption, and leasing c_2 . The excess of each sector is used by the rest of the economy, and inverse too. This *intersectoral* equilibrium must be maintained for the *simple reproduction*. This approach saves intimate relationship with the input-output which, in turn, amounts to pioneer classic analyzes.

A.1.2 The Extended Reproduction

The dynamics of a *growing economy*, that reproduces itself continuously, requires that the total gain with the *unequal trade national* are higher than the consumption needs of the system, part being accumulated to raise the amount of the means of production and to employ more *labor force*, which resets (or recalculates) the mass of *surplus-value* total or the *intersectoral equilibrium* equilibriums *asymmetrical* e^i , now added the portion \bar{s} :

$$e^i = s^I + s^{II} + \bar{s}$$

This amount e^i will be divide, on the one hand, for investments in sectors producing goods of production ($e_{1c}^i + e_{1v}^i$) and consumption ($e_{2c}^i + e_{2v}^i$) and, on the other hand, in refitting the merely necessary to economic reproduction simple, e_{r1}^i and e_{r2}^i , respectively, in *Sectors 1* and *2*. Emerges the following system:

$$\begin{aligned} P^N_1 &= c_1 + v_1 + e_{r1}^i + e_{1c}^i + e_{1v}^i \\ P^N_2 &= c_2 + v_2 + e_{r2}^i + e_{2c}^i + e_{2v}^i \\ P^N &= c + v + e_r^i + e_c^i + e_v^i \end{aligned}$$

To the demand for means of production is equal to the sum of the needs of refitting and expansion of both sectors and so that the demand for consumer goods is identical to the wage's fund, increased by the expansion of this, increased by surplus consumed in both sectors, the demands total of *means of production* (D_1) and consumption (D_2) composes the equations:

$$\begin{aligned} D_1 &= c_1 + c_2 + e_{1c}^i + e_{2c}^i \\ D_2 &= v_1 + v_2 + e_{1v}^i + e_{2v}^i + e_{r1}^i + e_{r2}^i \end{aligned}$$

The condition of equilibrium in the sector of the means of production ($P^N_1 = D_1$) is given:

$$c_1 + v_1 + e_{r1}^i + e_{1c}^i + e_{1v}^i = c_1 + c_2 + e_{1c}^i + e_{2c}^i$$

Simplifying, we have, in the same way as for *simple reproduction*, the *relation input-output* which, as we shall see, must be respected in *extended reproduction*: $v_1 + e_{r1}^i + e_{1v}^i = c_2 + e_{2c}^i$. The condition of equilibrium in the sector of means of consumption ($P^N_2 = D_2$) is given:

$$c_2 + v_2 + e_{r2}^i + e_{2c}^i + e_{2v}^i = v_1 + v_2 + e_{1v}^i + e_{2v}^i + e_{r1}^i + e_{r2}^i$$

Simplifying, emerges the same *relation Input-Output* guests in Sector 1: $c_2 + e_{2c}^i = v_1 + e_{1v}^i + e_{r1}^i$. Structuring the system of sectorial interdependencies in extended reproduction:

$$\begin{aligned}
P^N_1 &= c_1 + e_{1c}^i + v_1 + e_{r1}^i + e_{1v}^i \\
P^N_2 &= c_2 + e_{2c}^i + v_2 + e_{r2}^i + e_{2v}^i
\end{aligned}$$

This is the configuration of systemic markets oligopsonics of solid waste in Brazil. Part of the product of the Sector 1, whose value is equal to $c_1 + e_{1c}^i$, remains within the sector, for *refitting of the means of production consumed* the remainder, **in bold**, with a value equal to the sum $v_1 + e_{r1}^i + e_{1v}^i$, is replaced with $c_2 + e_{2c}^i$ from Sector 2 that, reciprocally, retains $v_2 + e_{r2}^i + e_{2v}^i$ internally, *to its final consumption*. The outcome is similar to the end of the analysis of the conditions of previous reproduction: the excess of each sector is the demanded by the rest of the economy, and inverse. For the *extended reproduction*, such intersectoral relationship of equilibrium must be maintained. Fair Trade is equilibrium in the exchange and in the system, but the reverse is not necessarily true. The system can be inequilibrium with unequal trade.

Thus are patents the conditions and the structural possibilities of systemic reproduction capitalist with the presence of monopsonic elements. The exercise of market power, coercion explicit or simulated and subsequent unequal trade do not generate disequilibrium itself, only transfers *income and/or product, or unequal gains*. The *equilibrium asymmetricalis* broken only with the oversights of relations crosscutting techniques.

A.2 W. Leontieff's Systems¹⁸

A. 2.1 Preliminary Considerations

To understand the interdependence of n sectors, the input-output matrices, in its static version, back to the question of determining the level of production that each one of sectors should reach to the demand of different products is satisfied accurately on the basis of *technical relations*.

To simplify the model, which can acquire any degree of complexity, considering the totality of methods/technologies productive, the heterogeneity of goods and joint production, three initial abstractions will be made: 1) each industry or sector produces only a homogeneous goods (or more than one, since they are produced in fixed proportion between themselves); 2) each business or industry uses a fixed ratio of inputs (combination of production factors); 3) the production in all industries is subject to constants incomes in a way that, if the total inputs vary in proportion k , the product varies in identic proportion. Made these assumptions, for the production of a unit of any j -th goods, the inputs required, in terms of i -th goods, are a constant amount, now given by the notation a_{ij} by which, therefore, the first subscribed (i) indicates the *input* and the second (j) the *product*. Also called *coefficient of input and product*, where it says, for exam-

¹⁸ "The economy input-output has become, since the publication of the pioneering works of Leontieff[...], one of the branches of huge importance to the economic sciences, due especially to its practical character support to the policies. In accordance with this development, the theory that underlies the use of matrices to model the economic structure of society has been concomitantly with the development of computers. If on the one hand, she allowed an economic modeling complexity unimaginable in past recent, this has led to the theory of programming and of economic planning is a very advanced to economists outside of this tradition, which devote themselves primarily to historical studies of the development of social relations that reflect the technical basis of production." (Lopes, Neder, 2011, p. 2).

ple, that $a_{ij} = 0,65$ already indicates that 0,65 units of goods i are required as input to produce one unit of product j , in general.

A. 2.2 Simple Reproduction Without Unequal Exchange

In this Table 1 are highlighted in gray the *coefficients of inputs* of products that each sector (I, II, III, ..., n) generates for itself, the other to meet the needs of the rest of the economy. Of course, some of these cells are null.

		Product					
Sectors		I	II	III	...	N	Total
Input	I	a_{11}	a_{12}	a_{13}	...	a_{1n}	X_1
	II	a_{21}	a_{22}	a_{23}	...	a_{2n}	X_2
	III	a_{31}	a_{32}	a_{33}	...	a_{3n}	X_3
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	N	a_{n1}	a_{n2}	a_{n3}	...	a_{nn}	X_n
	Total	Y_1	Y_2	Y_3	...	Y_n	VBP

Table 1: Sectorial Interflows

Source: Own Elaboration.

Should be noted that the *Gross Value of Production* (VBP) is the aggregate of any totals. This is still a closed model, where all demands are considered intermediate, the final demand does not exist. Emphasize, therefore, among n industries or sectors, a sector that determines a *demand non-intermediate*, but *final*, to product of each industry (d_1, d_2, \dots, d_n), the families, in return for the provision of a *primary input*, which is: the labor services. The *preliminary condition of equilibrium* is that, e.g., if the *Industry I* produces value enough to meet not only your needs, but those of all the other sectors in terms of their product (x_1), including the *final demand*, its production expressed an equation: $x_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n + d_1$. Isolating d_1 , hence reversing the signs and segregating x_1 in evidence, it follows that:

$$(1 - a_{11})x_1 - a_{12}x_2 - \dots - a_{1n}x_n = d_1$$

By doing the same procedure with the other sectors, will emerge this system:

$$\begin{aligned} (1 - a_{11})x_1 - a_{12}x_2 - \dots - a_{1n}x_n &= d_1 \\ -a_{21}x_1 + (1 - a_{22})x_2 - \dots - a_{2n}x_n &= d_2 \\ \dots & \\ -a_{n1}x_1 - a_{n2}x_2 - \dots + (1 - a_{nn})x_n &= d_n \end{aligned}$$

Translated into matrix notation, has algebraic operative opportunities:

$$\begin{bmatrix} (1 - a_{11}) & -a_{12} & \dots & -a_{1n} \\ -a_{21} & (1 - a_{22}) & \dots & -a_{2n} \\ \dots & \dots & \dots & \dots \\ -a_{n1} & -a_{n2} & \dots & (1 - a_{nn}) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix}$$

Denominating A the *Sectorial Interflows* of Table 1 we see that the first matrix of the equation above is equal to $-A$ added to the identity matrix (I) in which the elements of the first diagonal are always 1 and the others are zero. Hence:

$$(I - A)x = d$$

there x and d are, respectively, the vectors of variables *value* and *final demand*, and $(I - A)$ is usually called *technologic matrix*. Hence, by isolating the vector x , comes:

$$\bar{x} = (I - A)^{-1}d$$

This \bar{x} is the single solution, which expressed all *values of equilibrium* co-determined; it is sufficient that the technological matrix is *non-singular* (and there is no reason *a priori* to assume its *singularity*¹⁹) for its inverse $(I - A)^{-1}$ exists, and manages the \bar{x} .

A.2.3 Extended Reproduction and Unequal Exchange

Let us consider, therefore, the first equilibrium relationship, expressed in this equation:

$$X_i = \sum_{j=1}^n x_{ij} + x_i \tag{6}$$

where X_{ij} is the quantity of the product of the i -th sector transferred to the j -th, in which it is used as input, and x_i is the part of the product of this industry i -th not earmarked for other sectors, surplus, not appearing between the *relations of industrial interdependence*, and may x_i be *consumed*, *exported* or *accumulated*. It deduces the following system of *input-output*:

$$\begin{aligned} X_1 &= x_{11} + x_{12} \dots x_{1n} + x_1 \\ X_2 &= x_{21} + x_{22} \dots x_{2n} + x_2 \\ &\dots\dots\dots \\ X_n &= x_{n1} + x_{n2} \dots x_{nn} + x_n \end{aligned}$$

Can highlight among these lines the quantities of labor force employed in the production, indicating the whole by X_0 , which can be subdivided into the labor force employed in the generation of the products of i -ths sectors (X_{0i}), required them, highlighted in bold, and in the workforce used in addition to these needs (x_0). It follows an interflow sectorial:

$$\begin{aligned} X_0 &= \mathbf{x_{01}} + \mathbf{x_{02}} \dots \mathbf{x_{0n}} + x_0 \\ X_1 &= \mathbf{x_{11}} + \mathbf{x_{12}} \dots \mathbf{x_{1n}} + x_1 \\ X_2 &= \mathbf{x_{21}} + \mathbf{x_{22}} \dots \mathbf{x_{2n}} + x_2 \\ &\dots\dots\dots \\ X_n &= \mathbf{x_{n1}} + \mathbf{x_{n2}} \dots \mathbf{x_{nn}} + x_n \end{aligned}$$

¹⁹ That is to say, simply, that no row or column can have all its elements void, nor be linearly dependent, which would generate a determinate null too, and the vector prices would not have a single solution. But the assumption of sector zero, or with linear dependence of another, it is not realistic or likely.

Being this system conceived in terms of values, and not on physical quantities, it may be more appropriately called 'table of transactions'. The technological conditions of production are a synergy systemic continuous, can be described by those same *technical coefficients* or coefficients of production (a_{ij}), redefined as the reason of sector products i -th to the j -th (x_{ij}) and the total of productive demand of j -th (X_j), which implies: $a_{ij} = \frac{x_{ij}}{X_j} \therefore x_{ij} = a_{ij}X_j$. Replacing this definition x_{ij} set out in equation (6) emerges from this: $X_i = \sum_{j=0}^n a_{ij}X_j + x_i$. Considering $X_i = X_j$, can replace it in the above equation and generate this: $x_i = (1 - a_{ii})X_j - \sum_{j \neq i}^n a_{ij}X_j$. Distributing to all sectors, embodies the matrix of *technical coefficients*, identical to that which we have seen, equal to $-A$ increased the identity matrix (I), in form:

$$\begin{bmatrix} (1 - a_{11}) & -a_{12} & -\dots - & a_{1n} \\ -a_{21} & + (1 - a_{22}) & -\dots - & a_{2n} \\ \dots & \dots & \dots & \dots \\ -a_{n1} & -a_{n2} & -\dots + & (1 - a_{nn}) \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

The volume of production of inputs, including solid waste, which satisfies the needs of consumption crosscutting is:

$$\bar{X} = (I - A)^{-1}x$$

In the system are flagged absolute quantities, measured in physical units or values, but the system needs to be expressed in prices and quantities to come to the surface, how the transactions are showed to market players. Indique, therefore, the twofold schema notations:

- *Workforce:*

p_0 = Remuneration of labor force.

p'_0 = Remuneration of labor force employed in addition to the needs of refitting crosscutting;

q_0 = Quantity of workforce;

q'_0 = Quantity of employed work force in excess of the needs of refitting crosscutting;

- *Products of Work:*

$p_1, p_2 \dots p_n$ = Prices of the products in the sectors 1, 2, ..., n needed to interflow of economy;

p'_i = Price of the product of the sector i -th that exceeds the needs of interflow of economy.

$Q_i = \sum q_i = q_1 + q_2 + \dots + q_n$ = Total quantity of physical product of sector i -th;

q_i = Quantity of physical products of sectors 1, 2, ..., n necessary to interflow of economy;

q'_i = Quantity of physical product of sector i -th that exceeds the needs of interflow of economy.

Applied such prices at which quantities, has economic costs overall:

X_0 = Wage fund (total remuneration of labor);

$x_{0j} = p_0 q_{0j}$ = Wages necessarily employed for replacements crosscutting (Obs.: $p_0 q_{0j} = p_i q_{ij}$ when $i = 0$);

$x'_0 = p'_0 q'_0$ = Wage employed in addition to the needs of refitting crosscutting;

X_j = Aggregate cost of product of sector i -th;
 $x_{ij} = p_i q_{ij}$ = Cost of each product in the sector i -th to the j -th ($i=1, 2, \dots, n$).
 $x'_i = p'_i q'_i$ = Cost of the products of the sector i -th not intended to interflows sector;

Before structure in a matrix in orders these magnitudes, formalizes the two species of fundamental equations that constitute the model input-output:

$$X_0 = p_0 \sum_{j=1}^n q_{0j} + p'_0 q'_0 \quad (i = 0)$$

$$X_j = p_i \sum_{i=1}^n q_{ij} + p'_i q'_i \quad (j = 1, 2, \dots, n)$$

Expanding the summation or disaggregating each q_{ij} , emerges the matrix:

$$\begin{aligned}
 X_0 &= p_0 \sum q_{0j} + p'_0 q'_0 = \mathbf{p_0 q_{01}} + \mathbf{p_0 q_{02}} \dots \mathbf{p_0 q_{0n}} + p'_0 q'_0 \\
 X_1 &= p_1 \sum q_{1j} + p'_1 q'_1 = \mathbf{p_1 q_{11}} + \mathbf{p_1 q_{12}} \dots \mathbf{p_1 q_{1n}} + p'_1 q'_1 \\
 X_2 &= p_2 \sum q_{2j} + p'_2 q'_2 = \mathbf{p_2 q_{21}} + \mathbf{p_2 q_{22}} \dots \mathbf{p_2 q_{2n}} + p'_2 q'_2 \\
 &\dots\dots\dots \\
 X_n &= p_n \sum q_{nj} + p'_n q'_n = \mathbf{p_n q_{n1}} + \mathbf{p_n q_{n2}} \dots \mathbf{p_n q_{nn}} + p'_n q'_n
 \end{aligned}$$

$$\begin{array}{cccc}
 \mathbf{S_1} & + & \mathbf{S_2} & \dots & \mathbf{S_n} \\
 \mathbf{Y_1} & + & \mathbf{Y_2} & \dots & \mathbf{Y_n}
 \end{array}$$

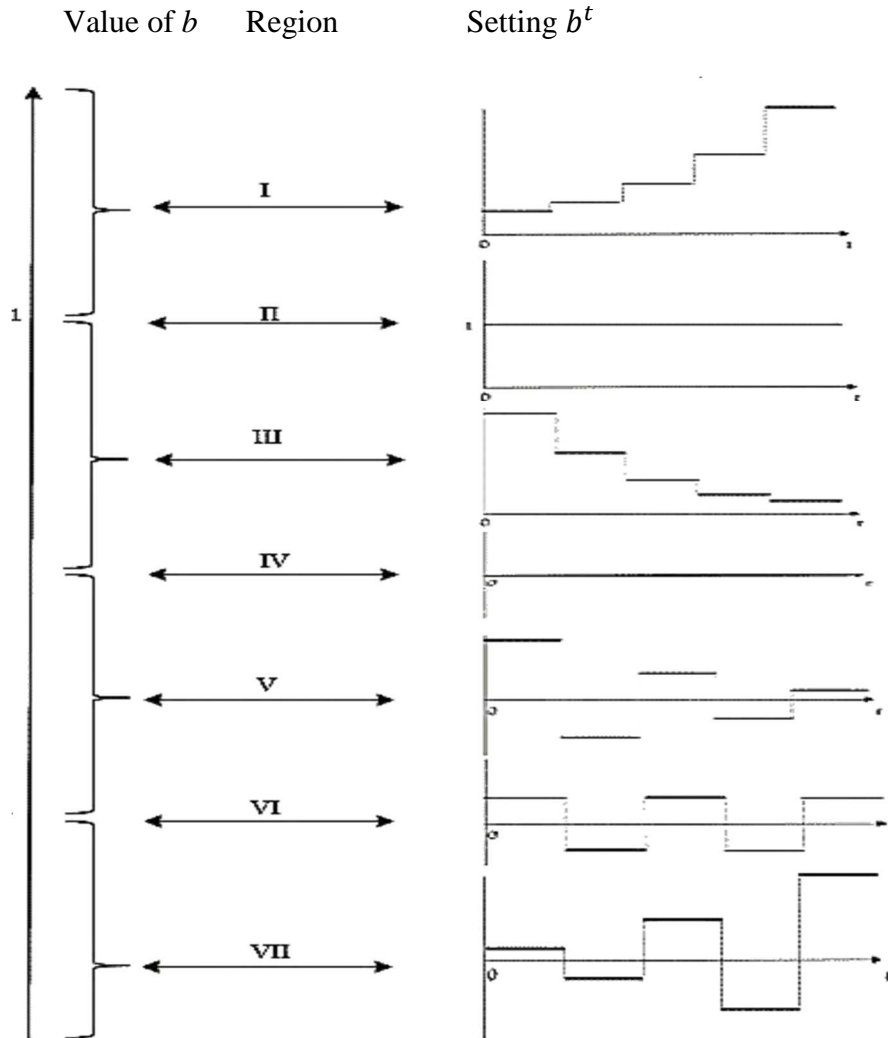
To demand for *refitting* is equivalent to offer it is necessary to reserve the last plots of horizontal sums (without bold), which does not maintain relationship with that demand, and on the basis of relations of interflow sectorial, not implying costs of production. The $\mathbf{p_i q_{ij}}$ ($i = 0, 1, 2, \dots, n$) are *horizontal plots* of inputs necessary to j -th products and total production costs incurred with the labor forces *and with the n products* : $x_{0j} q_{ij}$

$$\text{Equation of Costes} = x_{0j} + p_i \sum_{i=1}^n q_{ij} \quad (j = 1, 2, \dots, n) = X_j$$

The surplus S_j is the difference between the costs of the products of *sector* (X_j) and the sectorial prices market, that the classic called *cost of production* (Y_j) (with profit): $S_j = Y_j - X_j$ ($j = 1, 2, \dots, n$). Rewriting $Y_j = X_j + S_j$ and replacing X_j in the cost equation, extracts a theoretical relationship fundamental, identical to *Marxian decomposition of the value of the product* of a national economy (Lange, 1986, p. 66), given by:

$$Y_j = x_{0j} + \sum_{i=1}^n x_{ij} + S_j \quad (j = 1, 2, \dots, n)$$

- Temporal Trajectory of Ab^t Non-Oscillatory: if $b > 0$.
- Temporal Trajectory of Ab^t Oscillatory: if $b < 0$.
- Temporal Trajectory of Ab^t Explosive: if $|b| > 1$.
- Temporal Trajectory of Ab^t Compassionate: if $|b| < 1$



Graph 3: Dynamic Equilibrium of Non-Equilibrium
Source: (Chiang, 1982).

Obviously, this approach is abstract and must be econometrically adapted to local realities. Empirical studies show that the explosive behavior are not realistic, except in situations of crisis or exogenous shocks, insolvency of the offer or accelerated speculation. Rates of unequal trade usually converge to a swing point.

Now observe the solution of *Equation Non-Homogenous Complete*. Assuming the need for solution of an equation in differences of the first order, where a and c are constant, in the form $y_{t+1} + ay_t = c$, whose general resolution consists of the sum of two elements, which are: (I) a *particular integral* y_p (perhaps more appropriately called a *particular solution*, because it does not involve integration process) that resolves the *equation complete non-homogenous* above and (II) an additional function y_c , which is

the root of the equation $reducedy_{t+1} + ay_t = 0$, being that the component y_p represents the level of *equilibrium* of y and that, inversely, y_c responds by deflection of the *trajectory of equilibrium*. This element is the focus of analysis of the exercise of domination power. This sum solve the equation by induction of constant. Thus, the algebra will generates a solution from consider an initial condition of the waste market.

Employ, so evocative, for the *equation reduced* $y_{t+1} + ay_t = 0$ a solution in the form $y_t = Ab^t = y_c$, for which, it is logical, $y_{t+1} = Ab^{t+1}$ etc.

$$Ab^{t+1} + aAb^t = 0 \Leftrightarrow Ab^{t+1} = -aAb^t \Leftrightarrow b^{t+1} = -ab^t$$

The initial moment (when $t=0$) is defined: $b = -a \therefore y_c = A(-a)^t$. Alternative solutions, therefore, are in front of the analyst, initially. The simplest, of course, is the assumption that $y_t = k = y_p$, i.e., that y remains constant throughout the time, provided that it works, so that $y_{t+1} = y_t$, generating, by replacing in *equation non-homogenous complete*, these:

$$k + ak = c \quad e \quad k = \frac{c}{1+a} \therefore y_p(=k) = \frac{c}{1+a} \quad (a \neq -1)$$

In this case, the equilibrium would be stationary, provided that really $a \neq -1$, otherwise the solution would no longer be satisfactory, requiring new attempt of solution, not more assuming that $y_t = k$, but, for example, that $y_t = kt$, hence:

$$y_{t+1} = k(t+1) \therefore k(t+1) + akt = c \quad \text{and} \quad k = \frac{c}{t+1+at} = c \quad (\text{thus } a = -1)$$

Hence, $y_p(=kt) = ct$. Representing a function not constant of t , this second option points to a sort of equilibrium mobile, which, as far as the constant, just tell us about two forms of stability of the variables of an unequal exchange real. To determine it is necessary some tests, initially adding y_p and y_c and writing the *general solutions alternatives* in possible ways

$$y_t = A(-a)^t + \frac{c}{1+a} \quad (a \neq -1) \quad (F1)$$

$$y_t = A(-a)^t + ct \quad (a = -1) \quad (F2)$$

Being *anarbitrary* constant, both solutions are uncertain, still, being necessary, to $F1$, to use the *initial condition* ($t = 0$):

$$y_0 = A + \frac{c}{1+a} \Leftrightarrow A = y_0 - \frac{c}{1+a} \therefore y_t = \left(y_0 - \frac{c}{1+a}\right)(-a)^t + \frac{c}{1+a}$$

Also, in $F2$, $y_0 = A$, hence emerge the *solution* $y_t = y_0(-a)^t + ct$ or, remembering once again that, in this case, $a = -1$, we have the simply $y_t = y_0 + ct$.

ACKNOWLEDGMENTS:

First the Secretary of State for Labor, Employment, Income and Sports, Prof. PhD. Nilton Vasconcelos, for the opportunity to act in the management of the policy of Solidary Economy in Bahia, Brazil. The Superintendents of Solidary Economy of the Government of Bahia, Teach. PhD. Milton Barbosa Filho and Doc. Eng. Helbeth Lisboa de Oliva, by stimulating research. To Teach. PhD. Henrique Tomé and Teach. PhD. Vitor Athayde, by teaching and academic incentive. To economics Doc. Joselito Moura and Teach. MSCA. Plinio P. Moura, respectively, by the microeconomic debates and by some aspects of the analyses in classical political economy. To Teach. PhD. Rafael Antonio Sorhegui Ortega, for receiving at the University of Havana and the example of dedication to economic science. The Teach. PhD. Vilma Hidalgo y Teach. PhD. Yaima Doimadiós, for example to take on challenges in applied economics. The Teach. Doc. Jecilda Melo and the General Coordinator of AMACH, Mr. Cícero Melo, the partnership and institutional support.

Capítulo 6

EXPERIMENTOS DE SIMULACIÓN: ESTIMADOR POLINOMIAL LOCAL

Boukichou-Abdelkader, N.¹; Montero-Alonso, M.Á.²; Muñoz-García, A.³ y Canário, P.N.⁴

¹Centro de Investigación Ceiiis - IdiPAZ. Hospital Universitario La Paz. Madrid. España.

²Departamento de Estadística e Investigación Operativa, Universidad de Granada, España.

³Departamento de Estadística, Universidad Carlos III de Madrid, España.

⁴C3i, Polytechnic Institute of Portalegre, P -7300 -110, Portalegre, Portugal.

ABSTRACT

This simulation study estimates the trend of a nonparametric dataset using local polynomial estimators. The local regression technique is based on perform various adjustments parametric considering the near data to the point where you want to estimate the regression function. Necessarily, this simulation method is determined by three key parameters: the sample size, the type of domain or grid, where the data are simulated, and the model trend function. For the development of this technique, we used four R libraries: kernSmooth, locpol, locfit and sm, to practically implement the algorithms and interpret the results. Therefore, the objective of this simulated experiment is to facilitate a good adjusted model by applying the nonparametric estimate, using another alternative, when the trend model of observed data is unknown.

KEY WORDS: Nonparametric Regression, Local Polynomial Regression, type of Grid, sample size.

RESUMEN

En este estudio de simulación se pretende estimar la tendencia de un conjunto de datos de manera no paramétrica mediante el estimador polinomial local. La técnica de regresión local se basa en realizar varios ajustes paramétricos teniendo únicamente en cuenta los datos cercanos al punto donde se desea estimar la función de regresión. Previamente, para este método simulado es necesario determinar tres parámetros fundamentales: el tamaño muestral, el tipo de dominio o rejilla, donde se simularán los datos, y la función de tendencia para la simulación de los datos y la estimación del modelo. Para el desarrollo de esta técnica se utilizaran las librerías: kernSmooth, locpol, locfit y sm del software estadístico R, que facilitaran la aplicación práctica y la interpretación de los resultados simulados. Por tanto, el objetivo de este experimento simulado es facilitar un buen modelo ajustado aplicando la estimación no paramétrica, mediante otra alternativa, cuando se desconoce el modelo de tendencia de los datos observados.

Palabras claves: Regresión No Paramétrica, Regresión Polinomial Local, tipo de rejilla, tamaño muestral.

1. INTRODUCCIÓN

En la mayoría de las investigaciones científicas, uno de los problemas más importantes, es la formulación de los modelos estadísticos para representar de forma adecuada el fenómeno objeto de estudio.

En muchas ocasiones, no se dispone de toda la información real para realizar un análisis exhaustivo de los datos observados. Y en este momento es cuando se requiere la utilización de otros procedimientos más específicos para poder comprobar desde otra vía las sospechas que se tiene en la investigación objeto de estudio.

Afortunadamente, los contrastes paramétricos no son los únicos disponibles en la rama de la Estadística y existen otros métodos más flexibles para abordar este tipo de planteamiento, como son las técnicas no paramétricas. Las estimaciones no paramétricas permiten construir modelos que se ajustan a los datos de forma local ([1]), cuando no se puede asumir una distribución conocida.

Como es sabido, la teoría y los métodos de suavizamiento o regresión no paramétrica han cobrado un gran auge en las últimas décadas unido al avance en materia computacional, pudiéndose encontrar una revisión de los mismos en los libros [2], [3] y [4].

El creciente interés por estas metodologías ha tenido dos razones principales: la primera, que los planteamientos puramente paramétricos no aportaban la flexibilidad necesaria para la estimación de las curvas que aparecían en la práctica. Y la segunda razón, estaba ligada al avance de la informática y al desarrollo de un hardware que posibilitaba el costoso cálculo de esos estimadores no paramétricos.

Los primeros estimadores de regresión no paramétrica propuestos fueron los sencillos estimadores de tipo núcleo, [5] y [6].

Dichos estimadores se han ido modificando y moldeando dentro de los denominados **métodos de regresión polinomial local**, convirtiéndose en uno de los métodos más empleados por diversos analistas en la actualidad.

No obstante, para estos métodos de regresión no paramétrica se introduce uno de los problemas técnicos y críticos en la práctica, la elección del parámetro de suavizado o ancho de banda, que define la complejidad del modelo. Dada la dificultad que supone dicha selección, se distingue entre los métodos basados en la metodología plug-in y los basados en el criterio de validación cruzada.

En este sentido, el trato realizado de dichos métodos, ha sido dirigido fundamentalmente hacia la práctica, sin profundizar demasiado en aspectos teóricos de complejidad como son los procedimientos de tipo asintótico.

Desde este punto de vista no paramétrico, en concreto, con el estimador polinomial local, se pretende presentar un experimento simulado donde las observaciones analizadas son generadas mediante un modelo previamente definido, que está determinado aprioris por tres parámetros fundamentales, como son el tamaño muestral, el tipo de rejilla (o de dominio) donde se simularán los datos y la función de tendencia (tanto para la simulación de los datos como para la estimación de la tendencia).

Bajo este planteamiento se han explorado los métodos de regresión polinomial local como una de las mejores opciones de análisis, puesto que estas técnicas poseen características de flexibilidad, aplicabilidad e interpretabilidad bastante acertadas para este fin analítico. Para ello se ha realizado todo este proceso en el entorno de análisis y programación estadística R mediante algunas librerías específicas, como son *kernSmooth*, *locpol*, *locfit*, *ysm*, que se pueden descargar directamente a través de la web (<http://cran.es.r-project.org>).

2. MÉTODO

Sea un conjunto de n observaciones, $\{(X_i, Y_i), i=1, \dots, n\}$, buscamos un estimador de la función de regresión $m(x) = E[Y|X = x]$ de manera que los datos siguen el modelo,

$$Y_i = m(X_i) + \varepsilon_i \quad i = 1, \dots, n,$$

donde los residuos ε_i son variables aleatorias independientes con media cero y varianza $\sigma^2(X_i)$.

Para alcanzar tales objetivos se puede optar por una regresión paramétrica, y supone que la función de regresión desconocida, m , pertenece a alguna familia paramétrica de funciones, $m \in \{m_\theta | \theta \in \Theta\}$, donde θ se estima mediante mínimos cuadrados. La regresión no paramétrica no asume ninguna forma paramétrica para la función m ([1], [2], [3] y [4]), y la única restricción que se le impone es que sea suave, entendiendo esta suavidad en términos de derivabilidad.

Los primeros estimadores de regresión no paramétrica propuestos fueron los sencillos estimadores de tipo núcleo [5] y [6], estimadores que se han ido refinando y perfeccionando dentro de los denominados métodos de regresión polinomial local, convirtiéndose en uno de los métodos más empleados por diversos analistas en la actualidad, ya que obtiene un estimador sencillo y corrige de forma automática los efectos frontera.

La regresión polinomial local supone que la función de regresión m , tiene p derivadas en un punto x_0 , obteniéndose una aproximación para los valores en un entorno de x_0 .

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 + \dots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p,$$

es decir, se puede aproximar localmente m por funciones polinómicas de grado p

$$P_p(x) = \sum_{j=0}^p \beta_j (x - x_0)^j,$$

obteniéndose estimaciones de los coeficientes $\hat{\beta}_j$ con $j = 0, \dots, p$.

Con el fin de estimar m localmente mediante polinomios de grado p se considerara un problema de mínimos cuadrados ponderados:

$$\min \sum_{i=1}^n \left\{ Y_i - \sum_{j=0}^p \beta_j (X_i - x_0)^j \right\}^2 k_h(X_i - x_0)$$

donde h es un parámetro denominado ancho de banda o parámetro de suavizado que controla las observaciones que caen en cada entorno, $K_h(u) = h^{-1}K(\frac{u}{h})$, donde la función $K(\cdot)$, se denomina función núcleo. Dicha función define las ponderaciones que se asignan a cada observación en el entorno local considerado. Habitualmente se supone una densidad simétrica y con soporte compacto, y p es el grado del ajuste polinomiallocal¹.

3. EXPERIMENTOS DE SIMULACIÓN

Se ha realizado un análisis con datos simulados donde las observaciones analizadas son generadas mediante un modelo previamente definido y constituido por tres parámetros fundamentales: el tamaño muestral, el tipo de rejilla y la función de tendencia. Desde este punto de vista, se pretende ilustrar los métodos de regresión no paramétrica, en concreto, el estimador polinomial local (véase las citas de la [8] a la [16]).

El objetivo de este experimento será cuantificar la bondad de las estimaciones (dado que se conocen los modelos exactos) y además mostrar aspectos interesantes del problema de regresión como será el del efecto del tamaño muestral y la variabilidad de la muestra considerada.

Para estos aspectos, se estudiará el comportamiento de los estimadores con distintos tamaños de muestra ($n = 25, 50, 100$ y 500) y con distintas desviaciones típicas para los residuos del modelo ($0,3, 0,4$ y $0,1$). Tras esta comparativa, se pretende observar la convergencia de la curva teórica y asimismo ver cómo el problema de estimación se hace más difícil de resolver a medida que se va aumentando el valor de la desviación típica de los residuos del modelo.

Para realizar el planteamiento anterior se considerará el siguiente modelo de regresión:

$$Y = m(x) + \varepsilon \text{ donde } m(x) = \text{sen}(2x) + 2\exp(-16x^2)$$

donde x se genera según una distribución uniforme continua en el intervalo $(-2, 2)$ y los residuos se consideran normales con media 0 y desviación típica σ .

En primer lugar, se empezará comparando el estimador polinómico local (**EPL**) con grados $p = 0, 1, 3$. Fijando el parámetro ancho de banda en $h = 0,15$. Y en segundo

¹Dicho estimador queda determinado por tres parámetros, fijados o definidos a priori para este experimento: el ancho de banda, la función núcleo y el grado p .

lugar, se comparará el **EPL** con los distintos métodos de selección para el ancho de banda (plug-in, CV, regla del pulgar), fijando ajustes de grado $p = 1$.

Para cuantificar la precisión de las estimaciones resultantes se utilizará como criterio de error la suma residual de cuadrados sobre una rejilla de puntos de estimación. De este modo se evaluará el estimador sobre una red de puntos x_l $l=1,\dots,ngrid$, equiespaciada en $(-2, 2)$ de tamaño $ngrid = 500$. Una vez calculadas las estimaciones sobre la rejilla se calcularán los errores con la fórmula: $\frac{1}{500} \times \sum_{i=1}^{500} (m(x_i) - \hat{m}_h(x_i))^2$ y se compararan los resultados tomando la raíz cuadrada.

Bajo este diseño considerará la estimación con diferentes casos tomando distintos tamaños muestrales ($n= 25, 50, 100$ y 500) y $\sigma=0,4$. Para estos casos y tras implementar el código correspondiente, los resultados obtenidos se reflejan en conjunto en la siguiente **Figura 1**.

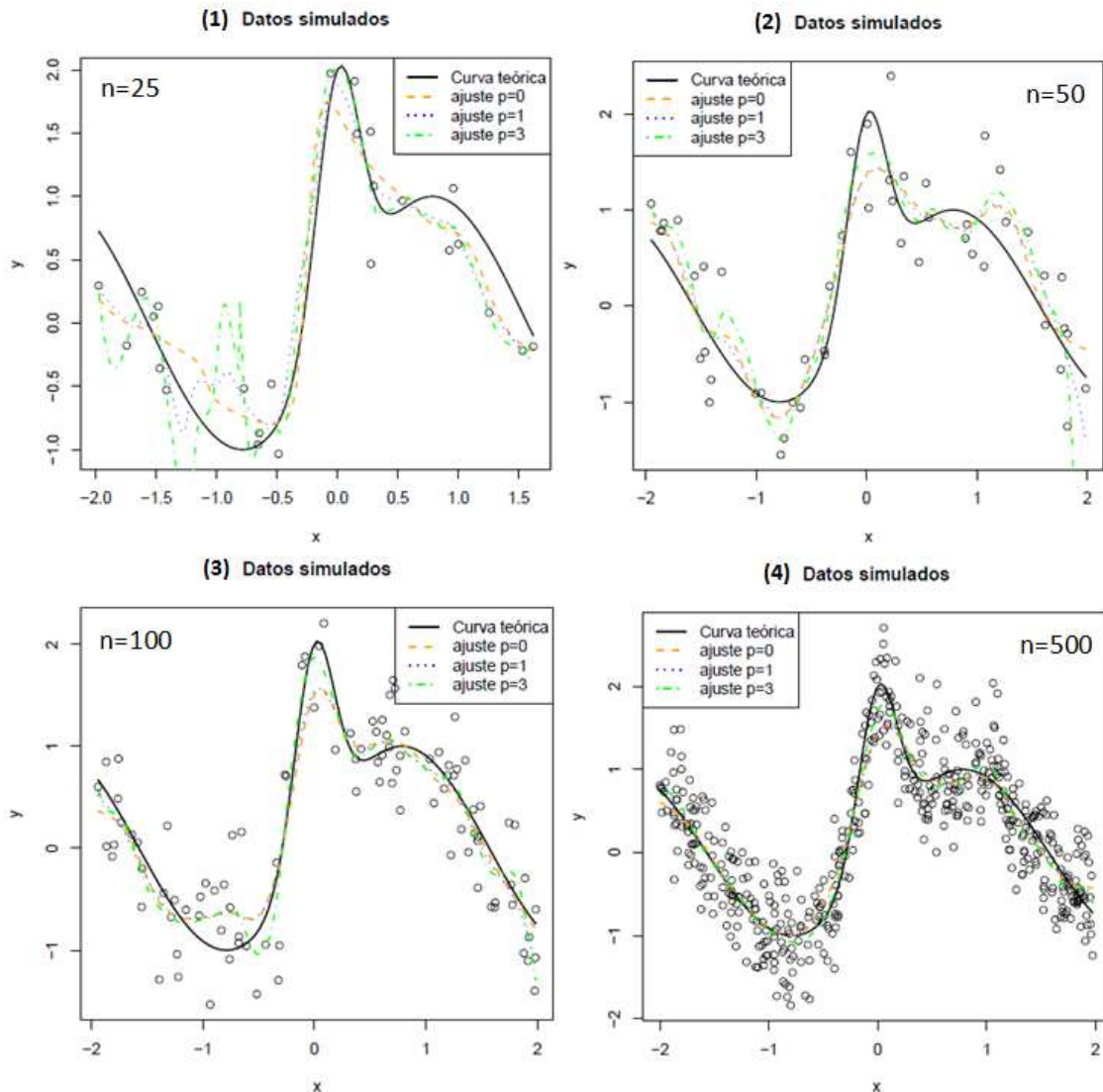


Figura 1: Estimación polinomial local a partir de datos simulados. Tamaños muestrales de 25, 50, 100 y 500 y desviación típica residual 0,4.

En este sentido, para el **caso 3**, como se puede observar en la **Figura 1**, el gráfico cuando $p = 0$ y $p = 1$ son muy parecidos los estimadores salvo en la frontera, debido a que $p = 1$ permite corregir de forma automática los efectos frontera. Los resultados para $p = 3$ muestran una mayor irregularidad.

Pero en el **caso 1**, **Figura 1**, cuando se ha disminuido el tamaño muestral a 25 se ve que los estimadores presentan bastantes irregularidades, sobre todo cuando se intenta ajustar un polinomio de grado alto ($p = 3$). Sin embargo, en el **caso 4**, **Figura 1**, cuando se considera un tamaño de muestra elevado se observa que los tres estimadores ajustados coinciden.

Por otro lado, también se ha ilustrado el comportamiento de los estimadores lineales locales con diferentes métodos de selección del parámetro de suavizado.

Esta aplicación comparativa se centra en el **caso 3**, que se simularon $n = 100$ datos con $\sigma = 0,4$, donde los métodos considerados son el selector basado en validación cruzada calculado usando la función **regCVBwSelC(cv)**, el de tipo plug-in calculado usando **pluginBw(pi)** y el calculado según la simple regla del pulgar, ofrecido por la función **thumbBw(th)**, todas ellas contenidas en la librería **locpol**. Y además, esta comparación de los métodos se realizará vía la raíz cuadrada del error definido anteriormente.

De esta manera, implementado el código correspondiente, los resultados obtenidos comparando con los tamaños muestrales $n = 25, 50, 100, 500$, según los diferentes selectores, son los siguientes:

	cv	th	pi
n= 25	0.3238807	0.2492679	NA
n= 50	0.2768001	0.2395432	0.2738346
n=100	0.2140991	0.2080944	0.2093302
n=500	0.07953835	0.09508265	0.08881742

De dichos resultados se puede observar que el comportamiento de los métodos plug-in es ligeramente superior a validación cruzada. No obstante la diferencia se hace menos patente en tamaños de muestra límite. También es de destacar que cuando se consideran pocos datos $n = 25$ no es posible el cálculo del selector de tipo plug-in. Esto es debido a que dichos métodos requieren estimaciones de las derivadas que no son posibles en este caso.

Y finalmente, se repite la comparación de los selectores variando la dificultad del problema de estimación. Esto se hará variando la desviación típica de los residuos del modelo, considerando $\sigma = 0.001, 0.1, 0.5$. Y el tamaño muestral se mantiene en $n = 100$. De este modo, los resultados obtenidos para los diferentes selectores son los siguientes:

	cv	th	pi
sigma= 0.001	0.01265973	0.08732214	0.03904589
sigma= 0.1	0.06306383	0.0912005	0.06364838
sigma= 0.5	0.2278124	0.1755693	0.1888333

Por tanto, la relación de los selectores cuando se muestran **distintos tamaños**

muestrales ($n=25, 50, 100, 500$) fijando la desviación típica ($\sigma = 0,4$) son muy diferentes, permitiendo reflejar que los métodos de **tipo *plug-in*** se comportan ligeramente superior a los de validación cruzada, aunque estas diferencias son menos patente en tamaños de muestra límite.

En contraposición, se detecta aun más la dificultad de la estimación cuando se modifican las **desviaciones típicas** de los residuos del modelo ($\sigma = 0.001, 0.1, 0.5$) y se fija el tamaño muestral en un valor en concreto ($n=100$), arrojando en los resultados obtenidos que cuando existe una **altavariabilidad muestral** ($\sigma = 0.5$) validación cruzada ofrece resultados poco representativos siendo el mejor ajuste el que viene mostrado por los selectores de **tipo *plug-in*** y en concreto, el de la sencilla regla del pulgar.

En resumen, con estos resultados obtenidos se ve un paralelismo entre aumentar el tamaño muestral y reducir la variabilidad residual. Si se observan estos datos cuando hay **alta variabilidad muestral** $\sigma = 0.5$ **validación cruzada ofrece resultados pobres**. Por eso, en este caso el **mejor ajuste** sería con los selectores de tipo *plug-in* y en particular con la sencilla regla del pulgar.

No obstante, nótese que a la vista de estos resultados aprioris se sabe que la **elección del ancho de banda** es uno de los aspectos cruciales del procedimiento de estimación para intentar buscar una adecuada **compensación entre sesgo y varianza**. Y además, se sabe que esta elección supone la complejidad del modelo.

También, esto mismo, ocurre con la **elección del grado de los ajustes polinomiales**. Es decir, la utilización de **ajustes de grado cero ó uno** nos darán estimaciones con **poca variabilidad**, muy suaves, pero con sesgos muy elevados. Por el contrario, cuando los ajustes son con **grados mayores** (dos ó tres) nos permitirán mayor adaptabilidad, o sea, menores sesgos pero, obtendremos **mayor varianza**.

Por eso, para este experimento se han fijado ajustes de grado $p = 1$, obteniendo de este modo varianzas pequeñas para poder obtener anchos de bandas aceptables que abarquen observaciones muy próximas al punto de estimación describiendo muy bien comportamientos locales y de esta manera poder comparar los resultados de los diferentes tipos de selectores calculados.

4. CONCLUSIONES

Con este experimento de simulación se pretende generar observaciones que puedan ser analizadas mediante el modelo previamente definido en base a los tres parámetros esenciales (el tamaño muestral, el tipo de rejilla y la función de tendencia).

El objetivo del experimento es cuantificar la bondad de las estimaciones, conociendo los modelos exactos, y además, mostrar aspectos interesantes del problema de regresión como son el del efecto del tamaño muestral y la variabilidad de la muestra considerada, permitiendo de esta forma visualizar los distintos comportamientos de los estimadores.

Además, desde esta comparativa, se pretende observar la convergencia de la curva teórica y asimismo ver cómo el problema de estimación se hace más difícil de resolver a

medida que se va aumentando el valor de la desviación típica de los residuos del modelo.

En esta aplicación se ha utilizado el software R como entorno de análisis y programación estadística y en concreto, algunas de las librerías específicas del mismo.

Los resultados logrados mediante la aplicación de estas técnicas no paramétricas sobre los datos simulados son:

- ✓ Los estimadores polinomiales con grados menores ($p=0$ y $p=1$) para una muestra de **tamaño 100** son muy parecidos salvo en la frontera, debido a que cuando $p = 1$ se corrigen de forma automática los efectos frontera. En cambio, con grados mayores ($p=3$) los resultados del estimador muestran demasiadas irregularidades en la curva.
- ✓ Cuando se disminuye el **tamaño muestral a 25**, los estimadores polinomiales presentan bastantes irregularidades sobre todo cuando se ajusta un polinomio de grado alto ($p = 3$).
- ✓ Los tres estimadores ajustados coinciden suficientemente bien sobre las observaciones cuando el tamaño muestral es bastante grande (**$n=500$**).
- ✓ El estimador polinomial local se comporta de forma distinta cuando se comparan varios tamaños muestrales y desviaciones típicas mediante la utilización de **diferentes métodos de selección** para el parámetro de suavizado reflejando resultados bastantes sorprendentes a los esperados.

En definitiva, se puede determinar que con estas confrontaciones realmente existe una semejanza entre aumentar el tamaño muestral y disminuir la variabilidad residual del modelo, y según los resultados recogidos, los métodos de tipo *plug-in* funcionan bastante mejor respecto a los de validación cruzada. También, se ha observado que la convergencia a la curva se alcanza cuando el tamaño muestral es bastante elevado ($n=500$) ya que, con tamaños menores se visualizan demasiadas irregularidades en la curva palpándose estas diferencias cuando se intentan ajustar polinomios de grado alto ($p=3$).

En esta línea exploratoria, una buena propuesta de investigación sería la aplicación de observaciones en **modelos aditivos** [17], cuya finalidad es la búsqueda de nuevos modelos ajustados que mejoren los resultados del estudio de una forma asequible para la rama de las ciencias sociales y de la biomedicina.

No obstante, para este experimento de simulación se puede concluir que los métodos de regresión no paramétrica, en concreto, el estimador polinomial local ofrecen una buena vía de solución e interpretación como fuente de análisis primaria, corroborando los resultados que aparecen en [7] y [18].

BIBLIOGRAFÍA

- [1] Boukichou-Abdelkader, N.; Montero-Alonso M. Á.; Muñoz-García, A. y Canário, P.N. (2014). **Regresión no paramétrica: estimador polinomial local**, En: Modelación matemática de fenómenos del medio ambiente y la salud (III), 46-52, Granada (España). ISBN: 84-616-7997-0.
- [2] Wand, M. P. and Jones, M. C. (1995). **Kernel Smoothing**. Chapman and Hall, London.
- [3] Fan, J. and Gijbels, I. (1996). **Local polynomial modelling and its applications**. Chapman and Hall, London.
- [4] Loader, C. (1999). **Local Regression and Likelihood**. Springer, New York.
- [5] Nadaraya, E.A (1964). On estimating regression. **Theory Probab. Appl**, 9, 141-142.
- [6] Watson, G. S. (1964). Smooth regression analysis. **Sankhya Serie A**, 26, 101-116.
- [7] Heckman, N.; Ramsay, J.O. (1996). **Spline smoothing with model based penalties**. McGill University, unpublished manuscript.
- [8] Ruppert, D.; Sheather, S. J.; Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. **Journal of the American Statistical Association**, 90, 1257-1270.
- [9] DeBrabanter, K.; De Brabanter, J.; De Moor, B. and Gijbels, I. (2013). **Derivative Estimation with Local Polynomial Fitting**. Journal of Machine Learning Research , 14, 281-301.
- [10] Tsybakov, A.B. (2009). **Introduction to Nonparametric Estimation**. Springer.
- [11] Zhou, S. and Wolfe, D.A. (2000). **On derivative estimation in spline regression**. Statist. Sinica, 10(1), 93-108.
- [12] Vilar-Fernández, J.A. and Vilar-Fernández, J.M. (1998). **Recursive estimation of regression functions by local polynomial fitting**. Ann. Ins. Stat. Math., 50(4), 729-754.
- [13] Francisco-Fernández, M. and Vilar-Fernández, J. M. (2001). Local polynomial regression estimation with correlated errors. Communications in Statistics: **Theory and Methods**, 30(7), 1271-1293.
- [14] Su, L. and Ullah, A. (2008). **Local polynomial estimation of nonparametric simultaneous equations models**. Journal of Econometrics, 144, 193-218.
- [15] Hall, P. and Yatchew, A. (2007). Nonparametric estimation when data on derivatives are available. **Annals of Statistics**, 35, 300-323.
- [16] Ichimura, H. and Todd, P. E. (2007). **Implementing Nonparametric and Semiparametric estimators**. Handbooks in Economics, 2, 5369-5468.
- [17] Hastie, T. J. and Tibshirani, R. (1990). **Generalized additive models**. Chapman and Hall. Washington, D.C.
- [18] Schimek, M. G. (2000). **Smoothing and regression: approaches, computation and application**. New York, Wiley.

Capítulo 7

ESTIMACIÓN DEL COSTO NIVELADO POR GENERACIÓN DE ENERGÍA ELÉCTRICA EN AEROGENERADORES: EVIDENCIA EMPÍRICA DE MÉXICO.

A. Santiago¹, J. M. Sautto², C.N. Bouza³ y Alfredo Montaña⁴

1. *Profesor de la Unidad Académica de Matemáticas de la UAG, e-mail: asantiago@uagro.mx*
2. *Coordinador del Campus Costa Chica de la UAG, e-mail: sautto1128@yahoo.com.mx;*
3. *Profesor de la Universidad de la Habana, Facultad de Matemática y computación. e-mail: bouza@matcom.uh.cu*
4. *Estudiante de la maestría en ciencias, área: estadística aplicada de la Unidad Académica de Matemáticas de la Universidad Autónoma de Guerrero.*

Resumen

A raíz de la reforma energética en nuestro país, se ha iniciado una discusión sobre las implicaciones positivas y negativas que esto puede tener en el ciudadano común y en los grupos interesados en invertir en generación de energía eléctrica y otras energías, usando lo que se ha denominado como energías limpias. En el presente trabajo se hace una valoración monetaria de la generación de energía eléctrica usando energía eólica y se estudia el caso de la generación de energía eléctrica en la zona del istmo de Tehuantepec, Oaxaca, México. Se presentan resultados sobre estimación de costos de generación, mediante distintos métodos de estimación, basados en la metodología de costos nivelados de producción de energía eléctrica. En el presente trabajo se han usado los métodos EPITRAG y el desarrollado por el CREE y, a partir de ellos se estima la utilidad neta que obtiene las empresas transnacionales interesadas en el tema.

Palabras clave: costo nivelado, generación eólica de electricidad, Tehuantepec.

1. Introducción.

1.1 El problema de investigación.

Ante la nueva reforma energética, se han elaborado distintas conjeturas en torno a los posibles beneficios, o inconvenientes, de abrir la generación de energías tradicionales y limpias en el país. La Secretaría de energía (SENER) y la actual administración federal, ha establecido como prioridad la transición energética rumbo a una generación y consumos más limpios¹. Para conseguirlo se ha promovido el desarrollo de un nuevo marco normativo que permita la difusión de información estratégica, así como la cooperación entre el Gobierno Federal y la iniciativa privada. Los beneficios que se espera obtener son mitigar la emisión de gases de efecto invernadero, contribuir a la

¹ “Prospectiva de Energías Renovables 2012-2026”, SENER, México.

diversificación de la matriz de generación de energía eléctrica, propiciando con esto un impacto positivo en la seguridad energética de México.

Hasta ahora, la generación eléctrica a partir de energías limpias en México a febrero del 2012, según datos de la CFE, la realizaban 204 centrales en operación y en construcción, principalmente en Oaxaca y Veracruz, con proyectos eólico y bioenergéticos, respectivamente. La producción global de electricidad en México, es de apenas 8.7%, muy distante de EUA (28.7%). Al 2026 se pretende aumentar la producción global de electricidad, utilizando energías limpias, al 35%, sobre todo porque México tiene una posición privilegiada en lo que se llama “cinturón solar” que lo ubica entre los principales países con mayor potencial solar en el mundo².

El presente trabajo de investigación, a partir del análisis de las distintas experiencias en este campo y de las posibles implicaciones que la nueva reforma energética tendrá sobre esta región de la costa del pacífico sur, con un potencial inexplorado de energía eólica, fotovoltaica, solar y minihidráulica, esperamos evaluar las experiencias actuales en el campo de las energías limpias, evaluar el impacto económico y social en la región de influencia e identificar las potencialidades de estos tipos de energía para la generación, principalmente, de energía eléctrica. Centramos nuestra atención en las dos primeras, considerando el costo nivelado de producción eléctrica a partir de diversos estimadores propuestos en la literatura, ver por ejemplo Alonso, G. et al. 2006, NRDC 2012, Escudero, J. M. 2008 y otros, que consideran estimadores deterministas. En el presente trabajo utilizamos tales estimadores proponiendo una estructura de modelo estadístico, estimando su media y varianza.

1.2 Objetivos del estudio.

1.2.1 General.

Evaluar las distintas experiencias de empresas de generación de energías limpias -eólica- en la región y su impacto en el desarrollo económico y social de la región.

1.2.2 Específicos.

- Estimar el costo nivelado de producción eléctrica para un aerogenerador, la utilidad bruta y neta
- Estimar, a partir de un estudio de opinión, el potencial impacto social que se genera en la región por la instalación de un sistema eólico.
- Estimar, a partir de un estudio de opinión, el potencial impacto económico que se genera en la región por la instalación de un sistema eólico.
- Estimar, a partir de un estudio de opinión, el potencial impacto al medio ambiente que se genera en la región por la instalación de un sistema eólico.
- Comparar los resultados del estudio de impacto con las estimaciones de costo de generación de energía eléctrica con un sistema eólico.

² SENER/AGM

2. Marco teórico conceptual.

La reforma energética abrió nuevas expectativas sobre el desarrollo económico y social en el país. Junto a estas expectativas se abrió también un debate de hasta donde abrir la inversión extranjera en el sector energéticos, sobre todo en lo que se refiere a las denominadas “energías limpias”. Por supuesto, como en toda reforma constitucional, existen voces opositoras y aquellas que se pronuncian a favor. Los impulsores de la reforma energética han planteado los objetivos siguientes:

- (i) Mejorar la economía de las familias: Bajarán los costos de los recibos de la luz y el gas. Al tener gas más barato se podrán producir fertilizantes de mejor precio, lo que resultará en alimentos más baratos.
- (ii) Aumentar la inversión y los empleos: Se crearán nuevos trabajos en los próximos años. Con las nuevas empresas y menores tarifas habrá cerca de medio millón de empleos más en este sexenio y 2 y medio millones más para 2025, en todo el país.
- (iii) Reforzar a Pemex y a CFE: Se le dará mayor libertad a cada empresa en sus decisiones para que se modernicen y den mejores resultados. Pemex y CFE seguirán siendo empresas 100% de los mexicanos y 100% públicas.
- (iv) Reforzar la rectoría del Estado como propietario del petróleo y gas, y como regulador de la industria petrolera.

La reforma energética también es una Reforma Verde, que favorecerá una mayor inversión en el desarrollo tecnológico y la adopción de fuentes de energía menos contaminantes y de bajo costo, como la solar, la eólica y el gas.

Algunos elementos vertidos en el debate son:

- Hoy en día los mexicanos pagamos tarifas eléctricas, en promedio, 73% más caras que en EEUU.
- En los últimos dos años hemos llegado a importar gas natural hasta en 19 dólares, de países como Nigeria y Qatar, cuando en México podríamos producirlo en 3 dólares.
- En la última década el desarrollo del gas natural en EEUU les ha permitido crear un millón setecientos mil empleos.
- En 2012, EEUU contaba con 137 pozos en aguas profundas del Golfo de México, mientras que nosotros solamente hemos perforado 6.
- En 2012, EEUU había otorgado más de 9,100 permisos para perforar pozos de gas natural en rocas de lutitas, mientras que en México perforamos solamente 3.
- Tras su proceso de Reforma Energética, Colombia logró reducir la tasa de desempleo de 17 a 11%.
- Entre 2003 y 2012 los ingresos petroleros en Colombia, como porcentaje del PIB, crecieron en 80%.
- Entre 2000 y 2012 los ingresos de Brasil por concepto de regalías petroleras se duplicaron.

- Estimaciones conservadoras indican que en México tenemos en el subsuelo más de 140 billones de pies cúbicos de gas natural, sin explorar, que aún no hemos comenzado a aprovechar. Esto equivale a más del doble de las reservas totales de gas natural del país.
- Cálculos conservadores indican que en aguas profundas tenemos 26 mil millones de barriles de petróleo crudo equivalente sin explorar.
- En suma la riqueza potencial que México tiene en campos de aguas profundas y lutitas es de 87 mil millones de barriles de petróleo crudo equivalente, lo que representa el 75% de los recursos prospectivos con los que cuenta el país.
- El gas natural es el energético fósil menos contaminante de bajo costo al que actualmente tenemos acceso.
- El gas natural es aproximadamente 30% menos contaminante que el combustóleo y 40% menos contaminante que el carbón.
- En términos de las emisiones de contaminantes a la atmósfera, sustituir combustóleo por gas natural en la generación de energía eléctrica equivale a retirar la mitad de los autos que actualmente circulan en la Ciudad de México, es decir 2.3 millones.
- De los 750 millones de toneladas de CO₂ que anualmente lanzamos a la atmósfera: el 25% proviene del transporte, el 17% de la industria y el 18% de la generación de electricidad.
- Se estima que en EEUU esta revolución genere ahorros en los próximos 3 años de mil dólares anuales por hogar.

Por otro lado, los opositores a las reformas afirman que se trata de un engaño, toda vez que es falso que se mejore la economía de las familias, al menos no en las regiones donde se impulsan proyectos de inversión cobijados por las reformas, y si hay derrama económica en las regiones es marginal, los empleos que se ofrecen a los residentes locales son los de menor calificación.

Dentro de este marco general, más allá de la disputa de hasta donde las reformas al marco legal y de la constitución, este estudio tiene como pretensión valorar el capital humano y social existente en la región, que posibilite un despegue económico regional. Establecer los impactos económicos y sociales que en materia de generación de energías limpias puede tener en la región más pobre y atrasada tecnológicamente del país, y establecer alternativas de políticas públicas que deban implementarse en la región con la finalidad de lograr los mejores impactos en el desarrollo económico sustentable y con una distribución justa de la riqueza que se genere; por supuesto, esto último escapa a nuestras manos porque corresponde a los tomadores de decisiones impulsar ese tipo de políticas.

3. Metodología

Se propone una metodología sustentada en tres ejes fundamentales. Primero un diagnóstico de gabinete sobre los distintos estudios realizados sobre las experiencias de generación de energías limpias en distintas partes del país con compañías extranjeras (españolas en el caso del Istmo de Tehuantepec) y los intentos fallidos en el Estado de Guerrero sobre energía fotovoltaica. En esta primera fase se determinarán las distintas

apreciaciones sobre estas experiencias y se analizarán las políticas públicas sobre las que se apoyaron.

Se sabe que cada tecnología usada para producir electricidad tiene características específicas, entre las que podemos mencionar: el tiempo de construcción, eficiencia, costo diferente de inversión, vida útil, etc. Debido a estas diferencias es muy difícil realizar una comparación entre tecnologías diferentes considerando solo una de estas características.

Debido a que pretendemos estudiar el impacto económico de cada una de estas tecnologías para generar energía eléctrica, debemos hacer una comparación de estas tecnologías entre sí y su contribución individual y en conjunto al desarrollo socio económico regional.

Para realizar comparaciones entre las diferentes tecnologías se utilizará la metodología del costo nivelado, la cual permite cuantificar el costo unitario de la electricidad (en kWh) generada durante toda la vida útil de la central. Como se trata de un valor promedio, esto posibilita la comparación inmediata de estos costos entre diferentes tipos de tecnologías de generación, en nuestro caso, la convencional, la fotovoltaica y la eólica.

La metodología de costo nivelado considera la energía eléctrica total que la planta producirá durante su vida útil y esta cantidad se divide entre el costo total de inversión (considerando la tasa de interés y el flujo de efectivo durante todo el tiempo de construcción) más los costos de generación, operación y mantenimiento, todo a valor presente.

Esta metodología toma en cuenta que el costo de instalación está afectado por la tasa de interés aplicada a la inversión y por el flujo de efectivo durante el tiempo de construcción.

El costo total nivelado es la suma del costo nivelado de inversión y el costo de producción. Este último es el generado durante la vida económica de la planta e incluye el costo nivelado de combustible y costo nivelado de operación y mantenimiento.

En la literatura podemos encontrar algunos procedimientos que nos ayudan en la estimación, según la presente metodología.

3.1 Costo nivelado de energía.

Según el Instituto Nacional de Investigaciones Nucleares (ININ) la metodología de costos nivelados, para el caso de energía nuclear, permite comparaciones entre diferentes tecnologías, cuantificar el costo unitario de la electricidad generada (el kWh) durante la vida de la central nuclear, permite una comparación inmediata con el costo de otras tecnologías alternativas y considera la generación total de electricidad durante toda la vida útil de la planta.

Esta metodología toma en cuenta que el costo de instalación (overnight cost) está afectado por la tasa de interés aplicada a la inversión y el flujo de efectivo durante el tiempo de construcción.

Por lo tanto, el Costo Total Nivelado (CTNG) es la suma del costo Nivelado de Inversión (CNI), que es el costo generado durante el tiempo de construcción y el costo de producción, generado durante la vida económica de la planta, que incluye el Costo Nivelado de combustible (CNC) y el Costo Nivelado de Operación y Mantenimiento (CNO&M).

Por lo anterior, el cálculo de los Costos Nivelados de Energía (Levelized Costs of Energy, LCoE o LEC) están definidos por la siguiente integral de tiempo en donde, para cada año “t”:

$$LEC = \frac{\sum_{t=1}^n \frac{I_t + M_t + F_t}{(1+r)^t}}{\sum_{t=1}^n \frac{E_t}{(1+r)^t}}$$

I_t = gastos de inversión en el año t, M_t = gastos de operación y mantenimiento, F_t = gastos de combustibles, E_t = electricidad generada, r = tasa de descuento y t = tiempo de vida de la planta.

Otras variables que se incorporan al cálculo del LCoE:

- La degradación prevista de la planta (fuente: información estadística del productor)
- La eficiencia de la misma (fuente: información técnica del productor)
- El costo de la entrega de energía (por ejemplo, la distancia de la planta del grid o del centro de consumo)
- El factor de capacidad (capacity factor) y potencia de la planta.
- El “peso” de los impuestos a lo largo del tiempo de vida (taxlife)

El resultado final (el LCoE) puede fuertemente modificarse en función de estas variables, además de depender directamente los 4 tipos de costos propios asociados a la planta: inversión inicial + mantenimiento + operación + dinero.

3.2 Otros métodos para el cálculo de costos nivelados.

Consideraremos en seguida otros métodos para el cálculo del costo nivelado.

Uno de los problemas más serios en los estudios de Factibilidad Técnica - Económica de Centrales Eólicas- está basado en la determinación exacta del cálculo del costo del kWh generado por dicha central. Para tal fin se detalla a continuación algunos de los métodos utilizados por las compañías de electricidad y la industria de la energía eólica para determinar el costo del kWh generado por un sistema de conversor de energía eólica.

3.2.1 Método Epritag.

El mismo es utilizado por el “Electric Power Research Institute”. Este método da un costo nivelado, considerando el valor presente de la inversión total y calcula los flujos de gastos en el tiempo, repartiendo los mismos en cuotas iguales todos los años. Es decir que convierte flujos de gastos desiguales en flujos constantes equivalentes. En forma simplificada tiene la expresión siguiente:

$$COE = FCR * \frac{CI}{8760CF} + OM$$

Donde:

COE = Costo de la energía generada en US\$/kWh

FCR = Factor o cargo fijo de gastos. Considera un valor presente de servicios de deuda, retorno de capital, tasas de interés e impuestos, y seguros.

CI = Costo de inversión. Siendo el monto total de la inversión llave en mano, expresado por unidad de potencia (US\$/kW).

CF = Factor de capacidad. Relación entre la energía promedio anual estimada y la que hubiera producido a la potencia nominal del aerogenerador.

OM = Costo directos de operación y mantenimiento por unidad de energía (US\$/kWh).

Pueden nombrarse las siguientes limitaciones a éste método:

- Asume la amortización de la deuda en un tiempo igual a la vida útil de la instalación.
- No permite considerar un retorno variable del monto de contado y del monto adeudado, como tampoco de los costos variables.

3.2.2 Método desarrollado por el Dr. Bastianón.

El método desarrollado por Bastianón es un método sencillo por el cual calcula en forma aproximada el costo de generación de un sistema conversor de energía eólica, y su expresión es:

$$CE = \frac{C}{Ea} \left[\frac{i}{1 - (1 + I)^{-N}} \right] + OM$$

Donde:

CE = Costo de producción en US\$/kWh

Ea = Energía media generada en un año en kWh.

i = Interés anual sobre el capital.

N = Años de vida útil de la instalación.

OM = Costos de operación y mantenimiento en US\$/kWh.

C = Costo de construcción en US\$.

Algunas limitaciones de este método son las siguientes:

- Asume la amortización de la deuda en un tiempo igual a la vida útil de la instalación.
- No permite considerar un retorno variable del monto de contado y del monto adeudado, como tampoco de los costos variables.

3.2.3 Método desarrollado por el CREE.

Este método, desarrollado por el Centro Regional de Energía Eólica (CREE), es utilizado en los estudios de factibilidad económica de centrales eólicas, tiene en cuenta el plazo de amortización de la instalación, siendo su expresión la siguiente:

$$COE = \frac{CI}{E} * (a + OM) = EIC * b$$

Donde:

COE = Costo de generación de corriente eléctrica en US\$/kWh.

CI = Costo total de inversión de la instalación, llave en mano, en US\$.

E = Estimación de la energía anual producida por un aerogenerador. Dicha estimación surge del modelo matemático MATWASP menos un 10%.

OM = Costos de operación y mantenimiento en US\$/kWh. Dichos costos surgen del valor del contrato anual de mantenimiento preventivo, realizado con la empresa proveedora del equipamiento; más el cálculo de repuestos necesarios anualmente estimado para un plazo de 20 años; más el valor del seguro anual. De no contar con dichos montos se estima en un 2% de CI.

a = Tasa de anualidad de los costos de capital.

EIC = Costos específicos de energía.

b = Costos anuales de servicio.

Esta fórmula toma en cuenta el plazo de amortización del equipamiento y la depreciación del mismo.

La tasa de anualidad puede calcularse de la siguiente forma:

$$a = \frac{q^n * (q - 1)}{q^n - 1}$$

Como puede observarse la tasa de anualidad depende del interés i , dado que $q = 1 + i$."

Cualquiera de los métodos mencionados puede utilizarse para la estimación del costo nivelado de generación eléctrica y, en general, pueden usarse para estimar el costo de producción de cualquier mercancía, siempre y cuando se cuente con la información que el método demanda.

Debido a lo anterior y en vista que existe poca información pública disponible sobre los costos de generación eléctrica y sobre la utilidad bruta y neta que se obtiene a partir de un sistema de generación eléctrica utilizando los métodos convencionales o los llamados "de energías limpias", en el presente trabajo solo se presentan las estimaciones obtenidas por dos métodos, el usado por el CREE y el método EPITRAG.

3.3 Estructura de los estimadores de costo nivelado.

Se revisó la estructura de los estimadores de costo nivelado de producción de energía eléctrica con objeto de hacer una comparación entre ellos. Se compararon el método Epritag, utilizado en sus estimaciones por "Electric Power Research Institute" y el método desarrollado por el Centro Regional de Energía Eólica (CREE) y que es utilizado en los estudios de factibilidad económica de centrales eólicas. Empezaremos por revisar la estructura de tales estimadores, en los cuales se observa que la mayoría de datos son constantes y solo CF es variable, toda vez que se trata de la proporción del tiempo que se mantiene en operación un aerogenerador durante el día y durante todo el año, toda vez que es el factor de capacidad de generación de energía eléctrica, que puede variar según la velocidad de los vientos durante el día y durante todos los días del año. Entonces, el estimador del total COE, se obtiene a partir de la expresión

$$COE = K * \frac{1}{CF} + OM, \text{ donde } K = FCR \left(\frac{CI}{8760} \right) \quad (1)$$

Donde, COE = Costo de la energía generada en US\$/kWh, FCR = Factor o cargo fijo de gastos. Considera un valor presente de servicios de deuda, retorno de capital, tasas de interés e impuestos, y seguros. CI = Costo de inversión. Siendo el monto total de la inversión llave en mano, expresado por unidad de potencia (US\$/kW). CF = Factor de capacidad. Relación entre la energía promedio anual estimada y la que hubiera producido a la potencia nominal del aerogenerador y OM = Costo directos de operación y mantenimiento por unidad de energía (US\$/kWh).

La forma original del estimador es

$$COE = FCR * \frac{CI}{8760CF} + OM \quad (2)$$

Lo cual muestra evidencia de que se trata de estimadores del tipo razón, por lo que, según lo establecido en la teoría, es posible estimar su media y varianza. Algunos procedimientos para estimar el valor esperado y varianza del estimador se pueden consultar, por ejemplo, en Mood (1974), Hansen, Hurwitz and Madow (1953), Sampath (2001), Fuller (2009), Särndal, et al. (2001) y Sarjinder Singh (2003).

Adaptando estos estimadores a nuestro caso, se obtiene que dado que los estimadores (1) y (2) son del tipo razón, aunque con una sola variable, que es una proporción $p = CF$, la media estará dada por

$$E \left[\frac{K}{Y} \right] \approx \frac{K}{\mu_Y} + \frac{K}{\mu_Y^3} \text{var}[Y] \approx K \left(\frac{1}{\mu_Y} + \frac{1}{\mu_Y^3} \text{var}[Y] \right)$$

Dado que $Y = CF = P$

$$E \left[\frac{K}{Y} \right] \approx \frac{K}{\mu_Y} + \frac{K}{\mu_Y^3} \text{var}[Y] \approx K \left(\frac{1}{P} + \frac{1}{P^3} PQ \right) \approx K \left(\frac{1}{P} + \frac{Q}{P^2} \right) \approx K \left(\frac{P + 1 - P}{P^2} \right) \approx \frac{K}{P^2}$$

y su estimador será

$$\widehat{E \left[\frac{K}{Y} \right]} \approx \frac{K}{\widehat{P}^2}$$

Donde $K = FCR \left(\frac{CI}{8760} \right)$, como se ha definido antes.

La varianza aproximada para una razón de variables aleatorias está dada por

$$\begin{aligned} \text{var} \left[\frac{X}{Y} \right] &\approx \frac{1}{\mu_Y^2} \text{var}[X] + \frac{\mu_X^2}{\mu_Y^4} \text{var}[Y] - \frac{\mu_X^2}{\mu_Y^6} (\text{var}[Y])^2 + \frac{1}{\mu_Y^4} (\text{cov}[X, Y])^2 \\ &\quad + 2 \frac{\mu_X}{\mu_Y^5} \text{var}[Y] \text{cov}[X, Y] \end{aligned}$$

y adaptando este resultados al caso que nos ocupa se obtiene

$$\text{var} \left[\frac{K}{Y} \right] \approx \frac{1}{P^4} PQ - \frac{1}{P^6} (PQ)^2 \approx \frac{Q}{P^3} - \frac{Q^2}{P^4} \approx \frac{PQ - Q^2}{P^4} = \frac{Q(P - Q)}{P^4}$$

$$V \left(K * \frac{1}{P} + OM \right) = K^2 V \left(\frac{1}{P} \right) \approx K^2 \left[\frac{V(P)}{P^2} \right] \quad (7)$$

Y su estimador

$$\widehat{V} \left(K * \frac{1}{P} + OM \right) = K^2 \widehat{V} \left(\frac{1}{P} \right) \approx K^2 \left[\frac{\widehat{V}(P)}{\widehat{P}^2} \right] \quad (8)$$

Es posible considerar el estimador para distintos diseños de muestreo, aunque eso se hará en trabajos posteriores.

4. Resultados y discusión.

Con la información disponible, ha sido posible aplicar los métodos EPRITAG y C.R.E.E. para obtener los costos nivelados para generación de energía eléctrica en los parques eólicos. La tabla 1 y 2 muestran los resultados.

Tabla 1: Costo nivelado de producción de energía eléctrica usando el método EPRITAG.

MW/h	COE (\$/kwh)	COE (us/kwh)	FCR (%)	CI (us/kwh)	CF	OM (us/kwh)
1	0.971716729	0.069408338	0.12	2000	0.439	0.007
2	0.534858364	0.038204169	0.12	1000	0.439	0.007
3	0.389235997	0.027802571	0.12	667	0.439	0.007

En esta tabla se observa que el costo de energía generada varía de \$0.97 (noventa y siete centavos MN) a \$0.39 (treinta y nueve centavos MN) por kwh generado, a medida que aumenta la capacidad del aerogenerador.

Estos costos incluyen la recuperación de la inversión incluyendo su tasa de interés, impuestos, seguros, mano de obra y demás gastos de administración.

A partir de estos cálculos podemos estimar la utilidad bruta por aerogenerador de 2000 kw de capacidad, por ejemplo. Sabemos que la tarifa de energía eléctrica para alumbrado público es de \$1.97 y que el F. P. de un aerogenerador es de 0.52 (para la zona del istmo de Tehuantepec), por lo que la producción anual será: $2000\text{kw} \times (365 \text{ días/año}) \times (24 \text{ horas/día}) \times (0.52) = 9,110,400 \text{ kw/año}$. Por lo anterior, la utilidad bruta obtenida por venta de energía eléctrica producida por una aerogenerador será: $9,110,400 \text{ kw/año} \times \$1.97/\text{kw} = \$17,947,488.00/\text{año}$. Ahora, si el costo de producción es de \$0.535/kw, entonces el costo de generación de energía eléctrica por año será de $9,110,400 \text{ kw/año} \times \$0.535/\text{kw} = \$4,874,064.00/\text{año}$ y la utilidad neta por aerogenerador será de: $\$13,073,424.00/\text{año}$.

Tabla 2: Costo nivelado de producción de energía eléctrica usando el método C.R.E.E.

MW/h	COE (\$/kwh)	COE (us/kwh)	CI	E	OM (us/kwh)	a
1	0.9901	0.0707	2,000,000	3,591,600	0.007	0.12
2	0.4950	0.0354	2,000,000	7,183,200	0.007	0.12
3	0.3300	0.0236	2,000,000	10,774,800	0.007	0.12

Usando este método se puede apreciar que las estimaciones de costos son similares a las obtenidas con el método EPRITAG, obteniendo de manera análoga los cálculos se encontró que:

Utilidad bruta = \$17,947,488.00/año. Costo de generación de energía eléctrica por año = \$4,509,648.60/año. Utilidad neta por un aerogenerador al año = \$13,437,840/año. Cálculos similares se pueden hacer para 1000kwh de capacidad y para 3000kwh. Los resultados se concentran en la tabla siguiente.

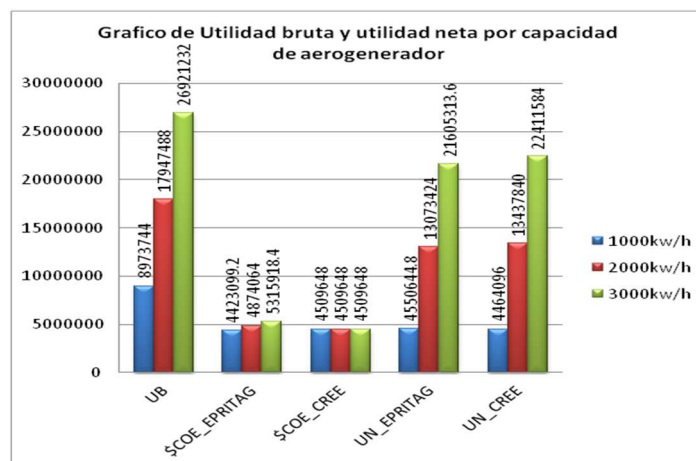
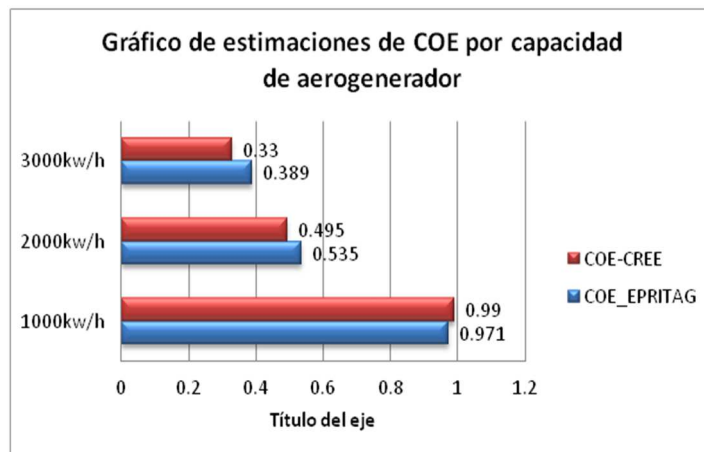
Tabla 3: Estimación de utilidad neta anual por capacidad de aerogenerador.

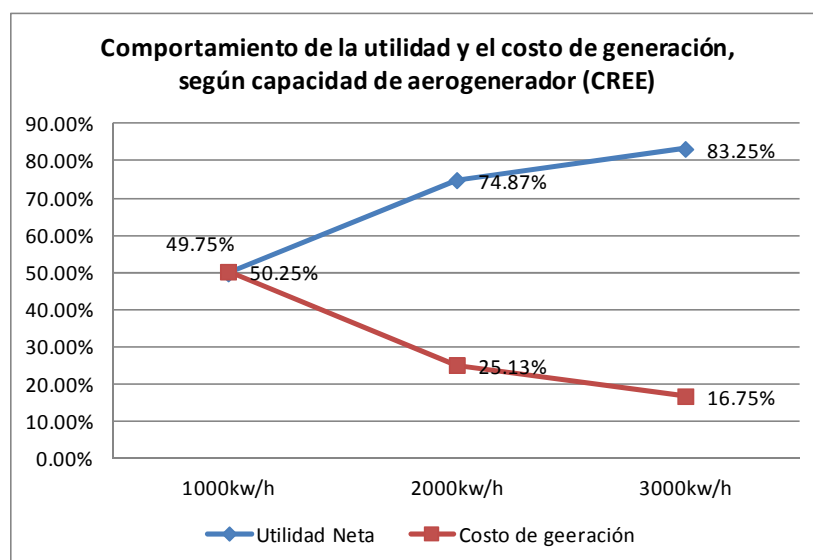
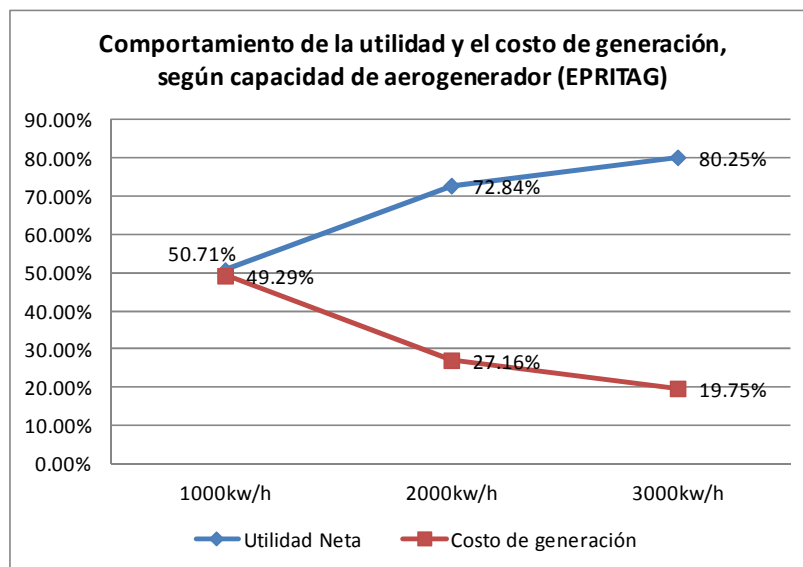
Generador	COE_EPRITAG	COE-CREE	UB	UCOE_EPRITAG	UCOE_CREE
1000kw/h	0.971	0.99	8973744	4423099.2	4509648
2000kw/h	0.535	0.495	1794748	4874064	4509648
3000kw/h	0.389	0.33	2692123	5315918.4	4509648

Tabla 4: Utilidad neta obtenida por aerogenerador según capacidad.

Generador	UN_EPRITAG	UN_CREE
1000kw/h	4550644.8	4464096
2000kw/h	13073424	13437840
3000kw/h	21605313.6	22411584

Donde, UB= Utilidad bruta por venta de energía a \$1.97, UCOE_EPRITAG = es la utilidad con el costo anual de producción calculado a partir de COE_EPRITAG, UCOE_CREE = es la utilidad con el costo anual de producción calculado a partir de COE_CREE. A partir de estos resultados definimos la utilidad neta para cada costo nivelado como $UN_EPRITAG = UB - UCOE_EPRITAG$ y $UN_CREE = UB - UCOE_CREE$.





5. Conclusiones.

Los resultados obtenidos muestran evidencia que el negocio de generación de energía eléctrica es altamente rentable para las empresas transnacionales que se han instalado en la zona del istmo de Tehuantepec. Para el ejemplo de cálculo considerado, que corresponde a una de las tarifas más bajas de venta de electricidad al público, el costo de producción eléctrica de un aerogenerador va de 25% al 27% de la utilidad bruta, por lo que la utilidad neta para la empresa será de 73% a 75% del total, para el caso de un aerogenerador de 2MGwh de capacidad.

La magnitud de este negocio seguramente nos da una idea del porqué de las dinámicas económicas, políticas y sociales en los distintos sectores interesados. Seguramente eso nos ayudará a explicar los cabildos para obtener concesiones, las prisas en la reforma energética, la criminalización de la lucha de grupos opositores, la cooptación de líderes sociales y políticos y, un largo etc.

Referencias.

- [1] Almonacid B. Ana, Nahuelhual M. Laura. *Estimación del potencial eólico y costos de producción de energía eólica en la costa de Valdivia, sur de Chile*. Agro Sur 37(2) 103-109: 2009
- [2] Alonso, G, Ramírez R. (2006). *Análisis de Costos Nivelados de la Generación de Electricidad en México*. Boletín energético 18.
- [3] Coviello M. *Energías Renovables en América Latina y el Caribe: buenas y malas noticias*. CEPAL-ONU
- [4] *Prospectiva del sector eléctrico 2012-2026*. SENER, Gobierno Federal. México.
- [5] *Programa Especial para el Aprovechamiento de Energías Renovables*. Diario Oficial. SENER, Gobierno Federal. México, 2009.

- [6] Almonacid B. Ana, Nahuelhual M. Laura. *Estimación del potencial eólico y costos de producción de energía eólica en la costa de Valdivia, sur de Chile*. Agro Sur 37(2) 103-109: 2009
- [7] Alonso, G, Ramírez R. (2006). *Análisis de Costos Nivelados de la Generación de Electricidad en México*. Boletín energético 18.
- [8] Coviello M. *Energías Renovables en América Latina y el Caribe: buenas y malas noticias*. CEPAL-ONU
- [9] Castillo Jara, Emiliano. *Problemática en torno a la construcción de parques eólicos en el istmo de Tehuantepec*. Revista Desarrollo Local Sostenible, vol. 4. No. 12.
- [10] Carta González, José Antonio; Calero Pérez, Roque; Colmenar Santos, Antonio y Manuel-Alonso Castro Gil. *Centrales de energías renovables*. Editoriales UNED y PEARSON EDUCACIÓN, S.A., Madrid, 2009.
- [11] *Prospectiva del sector eléctrico 2012-2026*. SENER, Gobierno Federal. México.
- [12] *Programa Especial para el Aprovechamiento de Energías Renovables*. Diario Oficial. SENER, Gobierno Federal. México, 2009.
- [13] Del Campo, M, Edelstein, N. *La energía del viento en México: Simulación de un parque eólico y aplicación de análisis probabilístico de seguridad*. RIIT, Vol. X. No.4. 2009.
- [14] *Programa de obras e inversiones del sector eléctrico (POISE)*. CFE, 2007-2016. México. 2006.
- [15] *Costos y parámetros de referencia para la formulación de proyectos de inversión. Generación 2005*. CFE, Subdirección de Programación. Gerencia de Evaluación y Programación de Inversiones. México. 2005.
- [16] Secretaría del Medio Ambiente. *Norma Oficial Mexicana. NOM-081-SEMARNAT-1994, Contaminación por Ruido*.
- [17] Henestroza OR. (2008). *Desarrollo del proyecto eólico en la región del istmo de Tehuantepec*. Investigación y Ciencia, uaa. 42: 18-21.
- [18] Elliot, D, et al. (2004). *Atlas de Recursos Eólicos del estado de Oaxaca*. Laboratorio Nacional de Energía Renovable. EUA, 138 pp.
- [19] Priego Cuevas, E. (2010). *Escenarios Económicos para la construcción de un nuevo reactor nuclear en México*. Tesis de maestría. Instituto Politécnico Nacional.
- [20] Palacios, J.C., Alonso, G., Ramirez, R., Gomez, A., Ortiz, J., & Longoria, L.C. (2004). *Levelized Costs for Nuclear, Gas and Coal for Electricity, under the Mexican Scenario*. Recuperado de <http://www.osti.gov/scitech/servlets/purl/840500>
- [21] Moreno Figueredo, C. (2007). *Diez preguntas y respuestas sobre energía eólica*. La Habana, Cuba: CUBASOLAR.

Capítulo 8

ESTIMACIÓN DE PARÁMETROS DE UN TEST DIAGNÓSTICO BINARIO BAJO UN MUESTREO TRANSVERSAL

José Antonio Roldán-Nofuentes

Bioestadística, Facultad de Medicina, Universidad de Granada, 18071, España

E-mail: jaroldan@ugr.es

Resumen

Un test diagnóstico es una prueba médica que se aplica a un individuo para determinar si tiene o no una cierta enfermedad, y su uso es fundamental en la práctica clínica. En este trabajo se realiza una revisión de los parámetros de un test diagnóstico binario. Se analizan las propiedades de los parámetros y se estiman estos parámetros cuando el test diagnóstico binario y el gold estándar se aplican a todos los individuos de una muestra aleatoria. Los resultados se han aplicado al diagnóstico de la enfermedad coronaria.

Palabras Clave: Coeficiente kappa ponderado, Coeficiente kappa promedio, Especificidad, Razones de verosimilitud, Sensibilidad, Test diagnóstico binario, Valores predictivos.

Abstract

A diagnostic test is a medical test that is applied to a patient in order to determine the presence or absence of a specific disease, and its use is essential in clinical practice. In this paper a review of the parameters of a binary diagnostic test is performed. The properties of the parameters are analysed and these parameters are estimated when the binary diagnostic test and the gold standard are applied to all of the individuals in a random sample. The results have been applied to the diagnosis of coronary artery disease.

Key Words: Weighted kappa coefficient, Average kappa coefficient, Specificity, Likelihood ratios, Sensitivity, Binary diagnostic test, Predictive values.

1. INTRODUCCIÓN

Un test diagnóstico es una prueba médica que se aplica a un paciente para determinar la presencia o ausencia de una cierta enfermedad. La mamografía para el diagnóstico del cáncer de mama o la ecocardiografía para el diagnóstico de la enfermedad coronaria son dos ejemplos de tests diagnósticos. En la práctica de la Medicina moderna el uso de los tests diagnósticos es fundamental, y aunque tal uso representa un bajo porcentaje en los costes en salud, sus resultados tienen una gran repercusión en la toma de decisiones médicas. Los tests diagnósticos se clasifican en:

- a). Binarios: dan lugar a dos posibles resultados, positivo (indica la presencia provisional de la enfermedad) o negativo (indica la ausencia provisional de la enfermedad). Ejemplo: ecocardiografía con dobutamina para el diagnóstico de la estenosis coronaria.
- b). Cuantitativos o continuos: dan lugar a valores numéricos. Ejemplo: concentración de glucosa en sangre en ayunas para el diagnóstico de la diabetes.

c). Ordinales: dan lugar a valores ordinales. Por ejemplo, una clasificación de la presencia de la enfermedad en definitivamente no, probablemente no, probablemente sí, definitivamente sí.

El problema que presentan los tests diagnósticos es que estos se pueden equivocar en el diagnóstico de la enfermedad, por lo que es necesario determinar su exactitud, la cual se expresa en términos de probabilidades o de funciones de probabilidades. Una revisión sobre la estimación de parámetros de tests diagnósticos se puede ver en los libros de Zhou [1] y de Pepe [2]. Para evaluar la exactitud de un test diagnóstico es necesario disponer de un estimador insesgado de dicha exactitud, para lo cual se necesita determinar el estado de enfermedad (enfermedad presente o enfermedad ausente) de cada individuo, independientemente del resultado del test diagnóstico. El procedimiento mediante el cual se conoce el verdadero estado de enfermedad de un individuo se denomina gold estándar. Por tanto, un gold estándar es una prueba médica que determina de forma objetiva si un individuo tiene o no una cierta enfermedad. Biopsia para el diagnóstico del cáncer de mama o la angiografía coronaria para el diagnóstico de la enfermedad coronaria son dos ejemplos de gold estándar.

Por tanto existen dos métodos que permiten diagnosticar la enfermedad: el test diagnóstico y el gold estándar. El test diagnóstico puede cometer errores en el diagnóstico de la enfermedad, mientras que el gold estándar no. Entonces, ¿por qué no utilizar siempre el gold estándar? Las siguientes razones justifican el uso de los tests diagnósticos:

1. El test diagnóstico suele ser más económico que el gold estándar.
2. El uso de un gold estándar puede suponer un riesgo para el individuo. Por ejemplo, la angiografía coronaria puede provocar ictus, trombosis e incluso la muerte del paciente.
3. No siempre existe un gold estándar. Por ejemplo, algunas enfermedades psiquiátricas.

El objetivo que se persigue cuando se evalúa la exactitud de un test diagnóstico no es determinar si un individuo tiene o no una cierta enfermedad, sino estimar su exactitud.

En este trabajo se realiza una revisión de parámetros de interés de un test diagnóstico binario (por ser de los más comunes en la práctica clínica y por su interés desde la perspectiva de los datos categóricos, pues su análisis da lugar al estudio de tablas 2×2 o de mayor dimensión). En la Sección 2 se describen los parámetros de interés. En la Sección 3 se estiman dichos parámetros bajo un muestreo transversal. En la Sección 4, se presenta un ejemplo real, y en la sección 5 se concluye.

2. PARÁMETROS DE UN TEST BINARIO

Considérese una población de individuos que pueden tener o no una cierta enfermedad. Supóngase que se dispone de un gold estándar para el diagnóstico de dicha enfermedad, entonces se define D como la variable aleatoria binaria que modeliza el resultado del GE, de tal forma que $D=1$ (gold estándar positivo) cuando el individuo tiene la enfermedad y $D=0$ (gold estándar negativo) cuando el individuo no tiene la enfermedad. La probabilidad de que un individuo (elegido al azar en la población objeto de estudio) tenga la enfermedad se denomina prevalencia de la enfermedad (p), esto es

$$p = P(D=1).$$

Considérese también un test diagnóstico binario cuya exactitud es evaluada con respecto a un gold estándar. Sea T la variable aleatoria que modeliza el resultado del test diagnóstico, de tal forma que $T = 1$ cuando el resultado del test es positivo (indicando la presencia provisional de la enfermedad) y $T = 0$ cuando el resultado del test es negativo (indicando la ausencia provisional de la enfermedad). A continuación se presentan parámetros del test diagnóstico binario.

2.1. Sensibilidad y especificidad

Los parámetros más importantes para evaluar la exactitud de un test binario son la sensibilidad (Se) y la especificidad (Sp). Ambos parámetros son las probabilidades condicionadas de clasificar correctamente a un individuo enfermo y a un individuo no enfermo respectivamente, y representan medidas de asociación o conformidad entre el test diagnóstico y el gold estándar.

La sensibilidad es la probabilidad de que el resultado del test binario sea positivo cuando el individuo está enfermo, es decir

$$Se = P(T = 1 | D = 1).$$

Se verifica que

$$Se + P(T = 0 | D = 1) = 1,$$

siendo $P(T = 0 | D = 1)$ la probabilidad de un falso positivo (pues estando el individuo enfermo, el test ha dado un resultado erróneo).

La especificidad (Sp) es la probabilidad de que el resultado del test binario sea negativo cuando el individuo no está enfermo, es decir

$$Sp = P(T = 0 | D = 0).$$

Se verifica que

$$Sp + P(T = 1 | D = 0) = 1,$$

siendo $P(T = 1 | D = 0)$ la probabilidad de un falso negativo (pues estando el individuo no enfermo, el test ha dado un resultado erróneo).

La sensibilidad y la especificidad del test binario dependen únicamente de la habilidad intrínseca del test diagnóstico para distinguir entre individuos enfermos e individuos no enfermos. Es decir, ambos parámetros dependen de las bases biológicas, físicas, químicas, etc..., con las que se ha desarrollado el test diagnóstico. Un test binario con una alta sensibilidad es útil para descartar la enfermedad ya que la probabilidad de un falso negativo es pequeña (y por tanto es útil para aplicar como procedimiento de rutina para el diagnóstico de la enfermedad). Un test binario con una alta especificidad es útil para confirmar la enfermedad ya que la probabilidad de un falso positivo es pequeña (y por tanto es útil para aplicarlo a individuos que son sospechosos de tener la enfermedad). A todo test binario hay que exigirle que

$$Se + Sp - 1 > 0.$$

Este parámetro ($Se + Sp - 1$) se denomina índice de Youden y sus valores varían entre -1 y 1. El índice de Youden presenta las siguientes propiedades:

a). Si el test binario no está relacionado con la enfermedad (aspecto que no es deseable), entonces la sensibilidad y la especificidad son complementarias y el índice de Youden es igual a cero. Esto es,

$$\text{si } P(T = i | D = i) = P(T = i) \Rightarrow \begin{cases} Se = P(T = 1) \\ Sp = P(T = 0) \end{cases} \Rightarrow Se + Sp = 1 \Rightarrow \\ Se = 1 - Sp \Rightarrow Y = 0.$$

b). Si el índice de Youden es menor que cero, entonces los resultados del test binario se deben intercambiar; esto es, $T = 1$ debe ser un resultado negativo y $T = 0$ debe ser un resultado positivo.

Una de las principales utilidades clínicas de un test diagnóstico es el screening. Un screening médico es un protocolo que se utiliza para detectar una enfermedad en individuos asintomáticos. El screening permite identificar individuos enfermos de forma temprana y su objetivo es reducir los efectos de la enfermedad en el individuo. Por ejemplo, screening del cáncer de mama, el screening del cáncer de próstata, etc... Para que un test diagnóstico se pueda utilizar como test de screening, se deben verificar las siguientes características:

1. Características de la población:
 - 1.1. La prevalencia de la enfermedad sea suficiente grande.
 - 1.2. Sea susceptible de aplicación de diferentes pruebas médicas y tratamientos.
2. Características de la enfermedad:
 - 2.1. Morbilidad y mortalidad significativas.
 - 2.2. Que tenga un tratamiento eficaz y aceptable.
 - 2.3. Periodo presintomático detectable.
 - 2.4. Mejora con un tratamiento precoz.
3. Características del test diagnóstico:
 - 3.1. Alta sensibilidad y especificidad.
 - 3.2. Bajo coste.
 - 3.3. Su aplicación suponga un bajo riesgo para el individuo.
 - 3.4. Exista un gold estándar.

2.2. Razones de verosimilitud

Para un resultado cada resultado del test binario, la razón de verosimilitud es un cociente de dos probabilidades definido como

$$LR = \frac{P(T = i | D = 1)}{P(T = i | D = 0)}, \quad i = 0, 1.$$

Si el resultado del test diagnóstico es positivo ($i = 1$), la razón de verosimilitud, denominada razón de verosimilitud positiva, es el cociente entre la sensibilidad y uno menos la especificidad, esto es

$$LR^+ = \frac{Se}{1 - Sp},$$

y representa el cociente entre la probabilidad de un resultado positivo del test diagnóstico en un individuo enfermo y la probabilidad de un resultado positivo del test diagnóstico en un individuo no enfermo. Cuando el resultado del test es negativo ($i = 0$), la razón de verosimilitud, denominada razón de verosimilitud negativa, es

$$LR^- = \frac{1 - Se}{Sp},$$

y representa el cociente entre la probabilidad de un resultado negativo del test diagnóstico en un individuo enfermo y la probabilidad de un resultado negativo del test diagnóstico en un individuo no enfermo. Las razones de verosimilitud solamente dependen de la sensibilidad y especificidad del test diagnóstico, y sus valores varían entre cero e infinito. Cuando el test binario y el gold estándar son independientes, entonces $LR^+ = LR^- = 1$. Si el test diagnóstico clasifica correctamente a todos los individuos (enfermos o no enfermos) entonces $LR^+ = \infty$ y $LR^- = 0$. Un valor $LR^+ > 1$ indica que un resultado positivo del test diagnóstico es más probable en un individuo enfermo que en un individuo no enfermo (como es de esperar), y un valor $LR^- < 1$ indica que un resultado negativo del test diagnóstico es más probable en un individuo no enfermo que en uno enfermo (como también es de esperar). Las razones de verosimilitud cuantifican el incremento sobre el conocimiento de la presencia de la enfermedad mediante la aplicación del test binario. Antes de aplicar el test binario, la odds de que un individuo tenga la enfermedad es

$$\text{pre-test odds} = \frac{p}{1 - p},$$

y después de aplicar el test, la odds de que un individuo tenga la enfermedad es

$$\text{post-test odds} = \frac{P(D = 1 | T = i)}{P(D = 0 | T = i)}, \quad i = 0, 1.$$

Por consiguiente, las razones de verosimilitud relacionan la pre-test odds con las post-tests odds, esto es

$$\text{Post-test odds } (T = 1) = LR^+ \times \text{pre-test odds}$$

$$\text{Post-test odds } (T = 0) = LR^- \times \text{pre-test odds},$$

y por tanto las razones de verosimilitud cuantifican el cambio en la odds de la enfermedad obtenido por el conocimiento de la aplicación del test diagnóstico binario.

2.3. Valores predictivos

La sensibilidad (especificidad) es la probabilidad de que el test diagnóstico dé un resultado positivo (negativo) cuando el individuo tiene (no tiene) la enfermedad. Sin embargo, cuando a un individuo se le aplica un test diagnóstico binario, el clínico o médico carece de información a priori sobre el verdadero estado de enfermedad (enfermedad presente o ausente) del individuo, por lo que desde el punto de vista práctico la sensibilidad y la especificidad no son parámetros que realmente interesen tanto al clínico como al individuo. Es decir, tanto el individuo como el clínico están

interesados en determinar cuál es el estado de enfermedad sabiendo el resultado del test diagnóstico. Los parámetros que aportan esta información son el valor predictivo positivo y el valor predictivo negativo, que son los parámetros que se utilizan para evaluar la exactitud clínica de un test binario. El valor predictivo positivo (*VPP*) es la probabilidad de que un individuo tenga la enfermedad cuando el resultado del test es positivo, y el valor predictivo negativo (*VPN*) es la probabilidad de que un individuo no tenga la enfermedad cuando el resultado del test es negativo. Los valores predictivos dependen de la sensibilidad y especificidad del test diagnóstico y también de la prevalencia de la enfermedad. Estos parámetros se calculan aplicando el Teorema de Bayes, esto es,

$$VPP = \frac{p \times Se}{p \times Se + (1-p) \times (1-Sp)} \quad \text{y} \quad VPN = \frac{(1-p) \times Sp}{p \times (1-Se) + (1-p) \times Sp}.$$

Los valores predictivos cuantifican el valor clínico del test diagnóstico, ya que tanto el individuo como el clínico están más interesados en conocer cómo de probable es tener la enfermedad dado un resultado del test diagnóstico. Un valor predictivo positivo alto indica que el test diagnóstico es útil para confirmar la enfermedad en la población cuya prevalencia de la enfermedad es p . Un valor predictivo negativo alto indica que el test diagnóstico es útil para descartar la enfermedad en la población cuya prevalencia de la enfermedad es p . El valor predictivo positivo es una función creciente de la prevalencia de la enfermedad, y el valor predictivo negativo es una función decreciente de la prevalencia.

2.4. Coeficiente kappa ponderado

Sean L y L' las pérdidas asociadas a una clasificación errónea con el test diagnóstico: L es la pérdida que se comete cuando en un individuo enfermo el test diagnóstico es negativo, y L' es la pérdida que se comete cuando en un individuo no enfermo el test diagnóstico es positivo. Las pérdidas L y L' son cero cuando un individuo (enfermo o no) es clasificado correctamente con el test diagnóstico. En la Tabla I se muestran las probabilidades y las pérdidas asociadas a la evaluación de un test diagnóstico binario con respecto a un gold estándar. En términos de las probabilidades y pérdidas de la Tabla I, la pérdida esperada al aplicar el test diagnóstico es

$$p(1-Se)L + q(1-Sp)L'$$

y la pérdida aleatoria es

$$p\{p(1-Se) + (1-p)Sp\}L + (1-p)\{pSe + (1-p)(1-Sp)\}L'.$$

La pérdida esperada es la pérdida promedio que se comete al clasificar erróneamente con el test diagnóstico a un individuo enfermo o no enfermo, y su rango de valores varía entre cero e infinito. La pérdida aleatoria es la pérdida que se comente cuando el test diagnóstico y el gold estándar son independientes, es decir cuando $P(T=i|D=j) = P(T=i)$. En términos de la pérdida esperada y de la pérdida aleatoria, el coeficiente kappa ponderado de un test binario [3] se define como

$$\kappa = \frac{\text{Pérdida aleatoria} - \text{Pérdida esperada}}{\text{Pérdida aleatoria}}$$

y es por tanto una medida del acuerdo más allá del azar entre el test diagnóstico y el gold estándar cuando ambos se aplican a una misma cohorte de individuos. Los valores del coeficiente kappa ponderado varían entre -1 y 1 . Sustituyendo en la ecuación anterior cada pérdida por su correspondiente expresión, se obtiene que el coeficiente kappa ponderado es

$$\kappa(c) = \frac{p(1-p)Y}{p(1-Q)c + (1-p)Q(1-c)},$$

donde $Q = pSe + (1-p)(1-Sp)$, $Y = Se + Sp - 1$ es el índice de Youden y $c = L/(L+L')$ es el índice de ponderación. Cuando la pérdida L es cero, entonces $c=0$ and el coeficiente kappa ponderado es

$$\kappa(0) = \frac{Sp - (1-Q)}{Q} = \frac{VPP - p}{1-p}.$$

Cuando la pérdida L' es cero, entonces $c=1$ y el coeficiente kappa ponderado es

$$\kappa(1) = \frac{Se - Q}{1-Q} = \frac{VPN - (1-p)}{p}.$$

Cuando $L=L'$, entonces $c=0.5$ y el coeficiente kappa ponderado (denominado coeficiente kappa de Cohen) es

$$\kappa(0.5) = \frac{p(1-p)Y}{\frac{p+Q}{2} - pQ} = \frac{2}{\frac{1}{\kappa(0)} + \frac{1}{\kappa(1)}} = \frac{2\kappa(0)\kappa(1)}{\kappa(0) + \kappa(1)},$$

por lo que el coeficiente kappa de Cohen es la media armónica de $\kappa(0)$ and $\kappa(1)$. Las expresiones $\kappa(0) = (Sp - (1-Q))/Q$ and $\kappa(1) = (Se - Q)/(1-Q)$ son la especificidad y la sensibilidad corregidas por azar según el modelo kappa, respectivamente.

El coeficiente kappa ponderado se puede escribir en términos de $p, Q, \kappa(0)$ y $\kappa(1)$ como

$$\kappa(c) = \frac{p(1-Q)c\kappa(1) + (1-p)Q(1-c)\kappa(0)}{p(1-Q)c + (1-p)Q(1-c)},$$

y por tanto el coeficiente kappa ponderado es una media ponderada de $\kappa(0)$ y $\kappa(1)$. También se puede escribir únicamente en términos de $\kappa(0)$ y $\kappa(1)$ como

$$\kappa(c) = \frac{\kappa(0)\kappa(1)}{c\kappa(0) + (1-c)\kappa(1)} = \frac{\varphi^2}{c\kappa(0) + (1-c)\kappa(1)}.$$

siendo $\varphi = \sqrt{\kappa(0)\kappa(1)}$ el coeficiente phi (coeficiente de correlación entre las variable aleatorias binarias T and D cuando estas se observan en todos los individuos).

El índice de ponderación c varía entre 0 y 1 y representa la pérdida relativa entre los falsos positivos y los falsos negativos. En la práctica el índice c es desconocido, pero sus valores se pueden intuir dependiendo del objetivo para el que se va a utilizar el test diagnóstico. Si el test diagnóstico se va a utilizar como paso previo a un tratamiento intensivo (es decir, como un test definitivo previo a un tratamiento de riesgo), hay más preocupación por los falsos positivos ($L' > L$) and $0 < c < 0.5$. Si el test diagnóstico se va a utilizar como un test de screening, hay una mayor preocupación por los falsos negativos ($L > L'$) y $0.5 < c < 1$. El índice c vale 0.5 cuando el test diagnóstico se utiliza para un diagnóstico simple (los falsos positivos y los falsos negativos tienen la misma importancia). El coeficiente kappa ponderado de un test binario presenta las siguientes propiedades:

- a). Si el acuerdo clasificatorio entre el test binario y el gold estándar es perfecto ($Se = Sp = 1$) entonces $\kappa(c) = 1$.
- b). Si la sensibilidad y la especificidad son complementarias ($Se = 1 - Sp$) entonces $\kappa(c) = 0$.
- c). Si la pérdida aleatoria es mayor que la pérdida esperada entonces $\kappa(c) > 0$; y si la pérdida aleatoria es menor que la pérdida esperada entonces $\kappa(c) < 0$ y los resultados del test diagnóstico deben intercambiarse ($T = 1$ debe ser el resultado negativo y $T = 0$ debe ser el resultado positivo). Por tanto, el análisis se debe limitar a los valores positivos del coeficiente kappa ponderado, y sus valores se pueden clasificar en la siguiente escala [10]: de 0 a 0.20 el acuerdo es malo, de 0.21 a 0.40 el acuerdo es mediocre, de 0.41 a 0.60 el acuerdo es moderado, de 0.61 a 0.80 el acuerdo es bueno, y de 0.81 a 1 el acuerdo clasificatorio es muy bueno o casi perfecto.
- d). El coeficiente kappa ponderado es una función del índice c que es creciente si $Q > p$, decreciente si $Q < p$, o constante igual al índice de Youden ($Y = Se + Sp - 1$) si $Q = p$.

2.5. Coeficiente kappa promedio

En la práctica las pérdidas L and L' definidas en la Sección 2.4 no se pueden determinar, por lo que el clínico suele asignar valores al índice de ponderación dependiendo de su conocimiento sobre la importancia relativa entre los falsos positivos y los falsos negativos. Así por ejemplo, si el clínico considera que los falsos positivos son el doble de importantes que los falsos negativos, entonces asignará al índice de ponderación c el valor $1/3$. Sin embargo, en muchas situaciones prácticas el clínico no dispone de un criterio o conocimiento que le permita fijar el valor del índice de ponderación. En esta situación, el clínico puede asignar distintos valores al índice de ponderación y analizar los resultados con cada uno de los valores fijados. Incluso en una misma situación práctica, diferentes clínicos pueden asignar valores distintos al índice de ponderación, dependiendo de sus propios conocimientos sobre el problema. Para resolver este problema de asignación de valores al índice de ponderación, Roldán-Nofuentes y Olvera-Porcel [4] han propuesto una nueva medida: el coeficiente kappa promedio.

Para unos valores fijos de sensibilidad, especificidad y prevalencia, el coeficiente kappa ponderado es una función continua del índice de ponderación c . Si el clínico considera que $L' > L$ (por ejemplo, si el test diagnóstico se va a utilizar como un test definitivo previo a un tratamiento de riesgo), entonces $0 < c < 0.5$ y el coeficiente kappa promedio se define como

$$\kappa_1 = \frac{1}{0.5} \int_0^{0.5} \kappa(c) dc$$

es decir, el coeficiente kappa promedio (κ_1) es el valor medio de la función $\kappa(c)$ cuando $0 < c < 0.5$. Si el clínico considera que $L > L'$ (por ejemplo, si el test diagnóstico se va a utilizar como un test de screening), entonces $0.5 < c < 1$ y el coeficiente kappa promedio se define como

$$\kappa_2 = \frac{1}{0.5} \int_{0.5}^1 \kappa(c) dc,$$

por lo que el coeficiente kappa promedio (κ_2) es el valor medio de la función $\kappa(c)$ cuando $0.5 < c < 1$. Resolviendo las integrales definidas se obtiene que

$$\kappa_1 = \begin{cases} \frac{2\kappa(0)\kappa(1)}{\kappa(0) - \kappa(1)} \ln \left[\frac{\kappa(0) + \kappa(1)}{2\kappa(1)} \right], & p \neq Q \\ Y, & p = Q, \end{cases}$$

y

$$\kappa_2 = \begin{cases} \frac{2\kappa(0)\kappa(1)}{\kappa(0) - \kappa(1)} \ln \left[\frac{2\kappa(0)}{\kappa(0) + \kappa(1)} \right], & p \neq Q \\ Y, & p = Q, \end{cases}$$

donde $\ln[\cdot]$ es el logaritmo neperiano. Cuando $p = Q$ el coeficiente kappa ponderado es siempre igual al índice de Youden (Y) para cualquier valor del índice de ponderación c , por lo que los dos coeficientes kappa promedio son también iguales al índice de Youden. Asimismo, para $p \neq Q$ los coeficientes kappa promedio se puede expresar en términos de $\kappa(0)$, $\kappa(1)$ and $\kappa(c)$ como

$$\kappa_1 = \frac{2[c\kappa(0) + (1-c)\kappa(1)]\kappa(c)}{\kappa(0) - \kappa(1)} \ln \left[\frac{\kappa(0) + \kappa(1)}{2\kappa(1)} \right]$$

y

$$\kappa_2 = \frac{2[c\kappa(0) + (1-c)\kappa(1)]\kappa(c)}{\kappa(0) - \kappa(1)} \ln \left[\frac{2\kappa(0)}{\kappa(0) + \kappa(1)} \right].$$

Como el coeficiente kappa ponderado es una medida del acuerdo más allá del azar entre el test diagnóstico y el gold estándar, los coeficientes kappa promedio (que se calculan a partir de coeficientes kappa ponderados) son medidas del acuerdo promedio más allá del azar entre el test diagnóstico y el gold estándar, y no dependen del índice de ponderación c . Los coeficientes kappa promedio κ_1 and κ_2 presentan las siguientes propiedades:

- a). Si $Se = Sp = 1$, entonces $\kappa_1 = \kappa_2 = 1$. Si $Se = 1 - Sp$, entonces $\kappa_1 = \kappa_2 = 0$. Por consiguiente, como la evaluación del test diagnóstico se debe limitar a los valores positivos del coeficiente kappa ponderado (propiedad (c) de la Sección 2.4), los valores de los coeficientes κ_1 and κ_2 son mayores que 0 and menores que 1.
- b). El coeficiente κ_1 es mayor que κ_2 si $p > Q$, and κ_1 es menor que κ_2 si $Q > p$.
- c). Para $p \neq Q$ las expresiones de κ_1 and κ_2 se pueden escribir como

$$\kappa_1 = 2\phi^2 \frac{\log[\kappa(0.5)] - \log[\kappa(0)]}{\kappa(1) - \kappa(0)}$$

y

$$\kappa_2 = 2\phi^2 \frac{\log[\kappa(1)] - \log[\kappa(0.5)]}{\kappa(1) - \kappa(0)},$$

respectivamente. Por tanto, el coeficiente kappa promedio κ_1 es proporcional al término $\frac{\log[\kappa(0.5)] - \log[\kappa(0)]}{\kappa(1) - \kappa(0)}$, que es el cociente entre la máxima diferencia

entre los coeficientes kappa ponderados (en logaritmos) cuando $L' > L$ y la máxima diferencia posible entre los coeficientes kappa ponderados. De forma similar, el coeficiente kappa promedio κ_2 es proporcional a $\frac{\log[\kappa(1)] - \log[\kappa(0.5)]}{\kappa(1) - \kappa(0)}$, que es

el cociente entre la máxima diferencia entre los coeficientes kappa ponderados (en logaritmos) cuando $L > L'$ y la máxima diferencia posible entre los coeficientes kappa ponderados.

d). El coeficiente kappa ponderado es una función continua en el intervalo $[0,1]$, por lo que el coeficiente kappa promedio κ_1 (κ_2) coincide con un valor del coeficiente kappa ponderado. De esta forma, una vez estimado el coeficiente kappa promedio, puede determinar el valor del índice de ponderación asociado al coeficiente kappa promedio estimado. Por tanto, la estimación del coeficiente kappa promedio permite estimar la pérdida relativa entre los falsos positivos y los falsos negativos asociada a este parámetro.

e). El coeficiente kappa promedio κ_1 minimiza la expresión $2 \int_0^{0.5} \{\kappa(c) - x\}^2 dc$. Cuando $x = \kappa_1$ esta expresión es la varianza del coeficiente kappa ponderado en torno a κ_1 . De forma similar, el coeficiente kappa ponderado κ_2 minimiza la expresión $2 \int_{0.5}^1 \{\kappa(c) - x\}^2 dc$. Cuando $x = \kappa_2$, esta expresión es la varianza del coeficiente kappa ponderado en torno a κ_2 .

3. ESTIMACIÓN DE LOS PARÁMETROS

La estimación de los parámetros se va a realizar bajo un muestreo transversal. Este tipo de muestreo consiste en aplicar el test diagnóstico binario y el gold estándar de forma independiente a cada uno de los individuos de una muestra aleatoria de tamaño n , y es

uno de los tipos de muestreo que más se aplican en la práctica. En la Tabla 1 se muestran las frecuencias observadas bajo un muestreo transversal.

Tabla 1. Frecuencias observadas bajo un muestreo transversal.

	$T = 1$	$T = 0$	Total
$D = 1$	s_1	s_0	s
$D = 0$	r_1	r_0	r
Total	$s_1 + r_1$	$s_0 + r_0$	n

3.1. Sensibilidad y especificidad

Condicionando en los totales de las filas, es decir en el resultado del gold estándar (variable D), se obtiene que la frecuencia s_1 es la realización de la distribución binomial $B(s, Se)$ y la frecuencia r_0 es la realización de la distribución binomial $B(r, Sp)$, por lo que los estimadores puntuales de la sensibilidad y de la especificidad son

$$\hat{Se} = \frac{s_1}{s} \quad \text{y} \quad \hat{Sp} = \frac{r_0}{r},$$

y son por tanto estimadores de proporciones binomiales, siendo sus varianzas estimadas

$$\text{Var}(\hat{Se}) = \frac{\hat{Se}(1-\hat{Se})}{s} \quad \text{y} \quad \text{Var}(\hat{Sp}) = \frac{\hat{Sp}(1-\hat{Sp})}{r}.$$

Un intervalo de confianza, denominado intervalo score modificado [5], para la sensibilidad es

$$Se \in 0.5 + \frac{s + \frac{z_{1-\alpha/2}^4}{s + z_{1-\alpha/2}^2}}{s + z_{1-\alpha/2}^2} (\hat{Se} - 0.5) \pm \frac{z_{1-\alpha/2}}{s + z_{1-\alpha/2}^2} \sqrt{\hat{Se}(1-\hat{Se})s + \frac{z_{1-\alpha/2}^2}{4}},$$

donde $z_{1-\alpha/2}$ es el $100(1-\alpha)\%$ percentil de la distribución normal estándar. Este intervalo es válido para $s \geq 10$. De forma similar, el intervalo de confianza score modificado para la especificidad es

$$Sp \in 0.5 + \frac{r + \frac{z_{1-\alpha/2}^4}{r + z_{1-\alpha/2}^2}}{r + z_{1-\alpha/2}^2} (\hat{Sp} - 0.5) \pm \frac{z_{1-\alpha/2}}{r + z_{1-\alpha/2}^2} \sqrt{\hat{Sp}(1-\hat{Sp})r + \frac{z_{1-\alpha/2}^2}{4}}$$

y es un intervalo válido para $r \geq 10$. El intervalo score modificado [5] es el intervalo de confianza para una proporción binomial que presenta un mejor rendimiento.

3.2. Razones de verosimilitud

Los estimadores máximo verosímiles de las LRs vienen dados por las expresiones

$$\hat{LR}^+ = \frac{s_1 r}{r_1 s} \quad \text{and} \quad \hat{LR}^- = \frac{s_0 r}{r_0 s}$$

y aplicando el método delta, sus varianzas asintóticas estimadas son

$$\widehat{Var}(\widehat{LR}^+) = \frac{\widehat{LR}^+}{(1 - \widehat{Sp})} \left(\frac{1 - \widehat{Se}}{s} + \frac{\widehat{Sp}}{r} \widehat{LR}^+ \right)$$

y

$$\widehat{Var}(\widehat{LR}^-) = \frac{\widehat{LR}^-}{\widehat{Sp}} \left(\frac{\widehat{Se}}{s} + \frac{1 - \widehat{Sp}}{r} \widehat{LR}^- \right).$$

Condicionando en los totales de filas de la Tabla I, las LR s son el ratio de dos proporciones binomiales independientes, por lo que las LR s se pueden estimar aplicando métodos para estimar el ratio de dos proporciones binomiales independientes. Martín-Andrés y Álvarez-Hernández [6] han estudiado intervalos de confianza para el ratio de dos proporciones binomiales independientes, recomendando utilizar los siguientes intervalos de confianza (adaptados a la notación de la Tabla I),

$$LR^+ \in \frac{n's_1'r_1' + \frac{z_{1-\alpha/2}^2}{2}(s's_1' + r'r_1' - 2s_1'r_1') \pm z_{1-\alpha/2} \sqrt{n'^2 s_1' r_1' (s_1' + r_1' - n'\hat{p}_1\hat{p}_2') + \frac{z_{1-\alpha/2}^2}{4}(s's_1' - r'r_1')^2}}{r_1' \left\{ n's_1'\hat{p}_1' - z_{1-\alpha/2}^2 (s' - r_1') \right\}}$$

y

$$LR^- \in \frac{n's_0'r_0' + \frac{z_{1-\alpha/2}^2}{2}(s's_0' + r'r_0' - 2s_0'r_0') \pm z_{1-\alpha/2} \sqrt{n'^2 s_0' r_0' (s_0' + r_0' - n'\hat{p}_3\hat{p}_4') + \frac{z_{1-\alpha/2}^2}{4}(s's_0' - r'r_0')^2}}{r_0' \left\{ n's_0'\hat{p}_3' - z_{1-\alpha/2}^2 (s' - r_0') \right\}}$$

donde $s_i' = s_i + 0.5$, $r_i' = r_i + 0.5$, $s' = s_1' + s_0'$, $r' = r_1' + r_0'$, $n' = s' + r'$, $\hat{p}_1' = r_1'/r'$, $\hat{p}_2' = s_1'/s'$, $\hat{p}_3' = r_0'/r'$ and $\hat{p}_4' = s_0'/s'$. Si el límite inferior del intervalo para LR^+ es menor que $s_1'/(n' - r_1')$ o mayor que el estimador de LR^+ , entonces el límite inferior del intervalo es

$$\frac{1}{s'(\hat{p}_1')^2 + z_{1-\alpha/2}^2} \left\{ s_1'\hat{p}_1' + \frac{z_{1-\alpha/2}^2}{2} - z_{1-\alpha/2} \sqrt{\frac{z_{1-\alpha/2}^2}{4} + s_1'(\hat{p}_1' - \hat{p}_2')} \right\},$$

y si el límite superior de este intervalo es mayor que $(n' - s_1')/r_1'$ o menor que el estimador de LR^+ , entonces el límite superior del intervalo es

$$\frac{1}{r'(\hat{p}_1')^2} \left\{ r_1'\hat{p}_2' + \frac{z_{1-\alpha/2}^2}{2} + z_{1-\alpha/2} \sqrt{\frac{z_{1-\alpha/2}^2}{4} + r_1'(\hat{p}_2' - \hat{p}_1')} \right\}.$$

Con respecto al CI para LR^- , si el límite inferior de este intervalo es menor que $s'_0 / (n' - r'_0)$ o mayor que el estimador de LR^- , entonces el límite inferior del intervalo es

$$\frac{1}{s'(\hat{p}'_3)^2 + z_{1-\alpha/2}^2} \left\{ s'_0 \hat{p}'_3 + \frac{z_{1-\alpha/2}^2}{2} - z_{1-\alpha/2} \sqrt{\frac{z_{1-\alpha/2}^2}{4} + s'_0 (\hat{p}'_3 - \hat{p}'_4)} \right\},$$

y si el límite superior de este intervalo es mayor que $(n' - s'_0) / r'_0$ o menor que el estimador de LR^- , entonces el límite superior del intervalo es

$$\frac{1}{r'(\hat{p}'_3)^2} \left\{ r'_0 \hat{p}'_4 + \frac{z_{1-\alpha/2}^2}{2} + z_{1-\alpha/2} \sqrt{\frac{z_{1-\alpha/2}^2}{4} + r'_0 (\hat{p}'_4 - \hat{p}'_3)} \right\}.$$

3.3. Valores predictivos

Condicionando en los totales de las columnas, es decir en el resultado del test diagnóstico binario (variable T), se obtiene que la frecuencia s_1 es la realización de la distribución binomial $B(s_1 + r_1, VPP)$ y la frecuencia r_0 es la realización de la distribución binomial $B(s_0 + r_0, VPN)$, por lo que los estimadores puntuales de los valores predictivos son

$$\hat{V}PP = \frac{s_1}{s_1 + r_1} \quad \text{y} \quad \hat{V}PN = \frac{r_0}{s_0 + r_0},$$

y son, al igual que la sensibilidad y la especificidad, estimadores de proporciones binomiales, siendo sus varianzas estimadas

$$\hat{V}ar(\hat{V}PP) = \frac{\hat{V}PP(1 - \hat{V}PP)}{s_1 + r_1} \quad \text{y} \quad \hat{V}ar(\hat{V}PN) = \frac{\hat{V}PN(1 - \hat{V}PN)}{s_0 + r_0}.$$

El intervalo score modificado [5] para el valor predictivo positivo es

$$VPP \in 0.5 + \frac{s_1 + r_1 + \frac{z_{1-\alpha/2}^4}{53}}{s_1 + r_1 + z_{1-\alpha/2}^2} (\hat{V}PP - 0.5) \pm \frac{z_{1-\alpha/2}}{s_1 + r_1 + z_{1-\alpha/2}^2} \sqrt{\hat{V}PP(1 - \hat{V}PP)(s_1 + r_1) + \frac{z_{1-\alpha/2}^2}{4}},$$

que es válido para $s_1 + r_1 \geq 10$. De forma similar, el intervalo de confianza score modificado [5] para el valor predictivo negativo es

$$VPN \in 0.5 + \frac{s_0 + r_0 + \frac{z_{1-\alpha/2}^4}{53}}{s_0 + r_0 + z_{1-\alpha/2}^2} (\hat{V}PN - 0.5) \pm \frac{z_{1-\alpha/2}}{s_0 + r_0 + z_{1-\alpha/2}^2} \sqrt{\hat{V}PN(1 - \hat{V}PN)(s_0 + r_0) + \frac{z_{1-\alpha/2}^2}{4}},$$

expresión válida para $s_0 + r_0 \geq 10$.

3.4. Coeficiente kappa ponderado

Sustituyendo en la expresión del coeficiente kappa ponderado cada parámetro por su estimador, esto es

$$\hat{S}_e = \frac{s_1}{s}, \quad \hat{S}_p = \frac{r_0}{r} \quad \text{y} \quad \hat{p} = \frac{s}{n},$$

el estimador máximo verosímil del coeficiente kappa ponderado es

$$\hat{\kappa}(c) = \frac{s_1 r_0 - s_0 r_1}{s(s_0 + r_0)c + r(s_1 + r_1)(1-c)},$$

con $0 < c < 1$. Aplicando el método delta la varianza asintótica estimada de $\hat{\kappa}(c)$ es

$$\hat{Var}(\hat{\kappa}(c)) = \frac{nr}{s \left[n^2(1-c)r_1 + n(cr_0 - 2(1-c)r_1)s_0 + n\{r_0 - (1-c)r_1\}s_1 + s(s_0r_1 - s_1r_0) \right]^4} \times$$

$$\left\{ (s_0r_1 - s_1r_0)^2 \left[2(1-c)r_1ns - (1-c)r_1n^2 + s(c(s_0r_0 + 2s_0r_1 + s_1r_1) - r_1s) \right]^2 + \right.$$

$$\left. s_1s_0nr^3 \left[(1-c)r_1n + s(cr - r_1) \right]^2 + r_1r_0nsr^2 \left[s_1r + c(s^2 - s_1n) \right]^2 \right\}.$$

Roldán Nofuentes et al[7] han estudiado diferentes intervalos de confianza para el coeficiente kappa ponderado. Dependiendo del tamaño muestral se pueden utilizar los siguientes intervalos de confianza:

a) Intervalo de confianza tipo Wald

Asumiendo la normalidad asintótica de $\hat{\kappa}(c)$, el intervalo de confianza tipo Wald para el coeficiente kappa ponderado es

$$\hat{\kappa}(c) \pm z_{1-\alpha/2} \sqrt{\hat{Var}(\hat{\kappa}(c))}.$$

Este intervalo de confianza tiene un buen rendimiento para muestras relativamente pequeñas ($n = 100$).

b) Intervalo de confianza bootstrap

El intervalo de confianza bootstrap se calcula generando K muestras con reemplazamiento de la muestra que se dispone y calculando en cada una de ellas el estimador del coeficiente kappa ponderado. Como estimador del coeficiente kappa ponderado se utiliza la media de las K réplicas de tales estimadores, y el intervalo de confianza global se calcula empleando el intervalo de confianza corregido por el sesgo [8]. Este intervalo de confianza tiene un rendimiento muy similar al intervalo de confianza tipo Wald.

c). Intervalo de confianza logit

Asumiendo la normalidad asintótica de $\hat{\kappa}(c)$, la transformación logit de $\hat{\kappa}(c)$, $\ln\left\{ \hat{\kappa}(c)/(1-\hat{\kappa}(c)) \right\}$, sigue una distribución normal de media $\ln\left\{ \kappa(c)/(1-\kappa(c)) \right\}$. De esta forma el intervalo de confianza para el logit de $\kappa(c)$ es

$$\text{logit}(\hat{\kappa}(c)) \pm z_{1-\alpha/2} \sqrt{\hat{V}ar(\text{logit}(\hat{\kappa}(c)))},$$

donde el estimador de la varianza del logit de $\hat{\kappa}(c)$ es

$$\hat{V}ar(\text{logit}(\hat{\kappa}(c))) = \frac{1}{[r_1 s - (1-c)r_1 n - c(s_0 r_0 + 2s_0 r_1 + s_1 r_1)]^2} \times$$

$$\left\{ \frac{s_1 s_0 r^2 [(1-c)r_1 n + (c r - r_1) s]^2 + r_1 r_0 s r [s_1 n - s_1 s + c(s^2 - s_1 n)]^2}{s(s_0 r_1 - s_1 r_0)^2} + \right.$$

$$\left. \frac{[2(1-c)r_1 n s - (1-c)r_1 n^2 + s(c(s_0 r_0 + 2s_0 r_1 + s_1 r_1) - r_1 s)]^2}{n s r} \right\}.$$

Finalmente, el intervalo de confianza logit para el coeficiente kappa ponderado es

$$\left(\frac{\exp\left\{\text{logit}(\hat{\kappa}(c)) - z_{1-\alpha/2} \sqrt{\hat{V}ar(\text{logit}(\hat{\kappa}(c)))}\right\}}{1 + \exp\left\{\text{logit}(\hat{\kappa}(c)) - z_{1-\alpha/2} \sqrt{\hat{V}ar(\text{logit}(\hat{\kappa}(c)))}\right\}}; \frac{\exp\left\{\text{logit}(\hat{\kappa}(c)) + z_{1-\alpha/2} \sqrt{\hat{V}ar(\text{logit}(\hat{\kappa}(c)))}\right\}}{1 + \exp\left\{\text{logit}(\hat{\kappa}(c)) + z_{1-\alpha/2} \sqrt{\hat{V}ar(\text{logit}(\hat{\kappa}(c)))}\right\}} \right),$$

y que tiene un buen rendimiento para muestras de tamaño igual o mayor a 200.

3.5. Coeficiente kappa promedio

Roldán-Nofuentes y Olvera-Porcel [4] han deducido los estimadores máximo verosímiles de los coeficientes kappa promedios y han estudiado varios intervalos de confianza para estos parámetros. Los estimadores máximo verosímiles de los coeficientes kappa promedio κ_1 and κ_2 son

$$\hat{\kappa}_1 = \begin{cases} \frac{2(s_1 r_0 - s_0 r_1)}{n_0 s - n_1 r} \log \left[\frac{n_1 r + n_0 s}{2n_1 r} \right], & s_0 \neq r_1 \\ \frac{s_1 r_0 - s_0 r_1}{s r}, & s_0 = r_1, \end{cases}$$

y

$$\hat{\kappa}_2 = \begin{cases} \frac{2(s_1 r_0 - s_0 r_1)}{n_0 s - n_1 r} \log \left[\frac{2n_0 s}{n_1 r + n_0 s} \right], & s_0 \neq r_1 \\ \frac{s_1 r_0 - s_0 r_1}{s r}, & s_0 = r_1, \end{cases}$$

respectivamente. En las ecuaciones anteriores $\hat{p} \neq \hat{Q}$ ($\hat{p} = \hat{Q}$) es equivalente a que $s_0 \neq r_1$ ($s_0 = r_1$). Asimismo, aplicando el método delta, cuando $s_0 \neq r_1$ las varianzas asintóticas estimadas de $\hat{\kappa}_1$ and $\hat{\kappa}_2$ son

$$\begin{aligned} \bar{\text{Var}}(\hat{\kappa}_1) &= \frac{1}{[\hat{\kappa}(0) + \hat{\kappa}(1)]^2 [\hat{\kappa}(0) - \hat{\kappa}(1)]^2} \times \\ &\left\{ \left[\frac{2\hat{\kappa}(0)^2 \hat{\kappa}(1) - \hat{\kappa}(1)[\hat{\kappa}(0) + \hat{\kappa}(1)]\hat{\kappa}_1}{\hat{\kappa}(0)} \right]^2 \frac{(1 - \hat{S}p)^2 \hat{Y}^2 \text{Var}(\hat{p}) + \hat{p}^2 [(1 - \hat{S}p)\text{Var}(\hat{S}e) + \hat{S}e^2 \text{Var}(\hat{S}p)]}{\hat{Q}^4} + \right. \\ &\left. \left[\frac{\hat{\kappa}(0)[(\hat{\kappa}(0) + \hat{\kappa}(1))\hat{\kappa}_1 - 2\hat{\kappa}(0)\hat{\kappa}(1)]}{\hat{\kappa}(1)} \right]^2 \frac{(1 - \hat{S}e)^2 \hat{Y}^2 \text{Var}(\hat{p}) + (1 - \hat{p})^2 [\hat{S}p\text{Var}(\hat{S}e) + (1 - \hat{S}e)^2 \text{Var}(\hat{S}p)]}{(1 - \hat{Q})^4} + \right. \\ &2 \left\{ \frac{2\hat{\kappa}(0)^2 \hat{\kappa}(1) + \hat{\kappa}(1)[\hat{\kappa}(0) + \hat{\kappa}(1)]\hat{\kappa}_1}{\hat{\kappa}(0)} \right\} \left\{ \frac{\hat{\kappa}(0)[(\hat{\kappa}(0) - \hat{\kappa}(1))\hat{\kappa}_1 - 2\hat{\kappa}(0)\hat{\kappa}(1)]}{\hat{\kappa}(1)} \right\} \times \\ &\left. \frac{(1 - \hat{p})\hat{p} [(1 - \hat{S}e)\hat{S}e\text{Var}(\hat{S}p) + (1 - \hat{S}p)\hat{S}p\text{Var}(\hat{S}e)] - (1 - \hat{S}e)(1 - \hat{S}p)\hat{Y}^2 \text{Var}(\hat{p})}{\hat{Q}^2 (1 - \hat{Q})^2} \right\} \end{aligned}$$

y

$$\begin{aligned} \bar{\text{Var}}(\hat{\kappa}_2) &= \frac{1}{[\hat{\kappa}(0) + \hat{\kappa}(1)]^2 [\hat{\kappa}(0) - \hat{\kappa}(1)]^2} \times \\ &\left\{ \left[\frac{\hat{\kappa}(1)[2\hat{\kappa}(0)\hat{\kappa}(1) - (\hat{\kappa}(0) + \hat{\kappa}(1))\hat{\kappa}_2]}{\hat{\kappa}(0)} \right]^2 \frac{(1 - \hat{S}p)^2 \hat{Y}^2 \text{Var}(\hat{p}) + \hat{p}^2 [(1 - \hat{S}p)\text{Var}(\hat{S}e) + \hat{S}e^2 \text{Var}(\hat{S}p)]}{\hat{Q}^4} + \right. \\ &\left. \left[\frac{\hat{\kappa}(0)[\hat{\kappa}(0) + \hat{\kappa}(1)]\hat{\kappa}_2 - 2\hat{\kappa}(0)\hat{\kappa}(1)^2}{\hat{\kappa}(1)} \right]^2 \frac{(1 - \hat{S}e)^2 \hat{Y}^2 \text{Var}(\hat{p}) + (1 - \hat{p})^2 [\hat{S}p\text{Var}(\hat{S}e) + (1 - \hat{S}e)^2 \text{Var}(\hat{S}p)]}{(1 - \hat{Q})^4} + \right. \\ &2 \left\{ \frac{\hat{\kappa}(1)[2\hat{\kappa}(0)\hat{\kappa}(1) - (\hat{\kappa}(0) + \hat{\kappa}(1))\hat{\kappa}_2]}{\hat{\kappa}(0)} \right\} \left\{ \frac{\hat{\kappa}(0)[\hat{\kappa}(0) + \hat{\kappa}(1)]\hat{\kappa}_2 - 2\hat{\kappa}(0)\hat{\kappa}(1)^2}{\hat{\kappa}(1)} \right\} \times \\ &\left. \frac{(1 - \hat{p})\hat{p} [(1 - \hat{S}e)\hat{S}e\text{Var}(\hat{S}p) + (1 - \hat{S}p)\hat{S}p\text{Var}(\hat{S}e)] - (1 - \hat{S}e)(1 - \hat{S}p)\hat{Y}^2 \text{Var}(\hat{p})}{\hat{Q}^2 (1 - \hat{Q})^2} \right\}, \end{aligned}$$

respectivamente, siendo $\hat{S}e = s_1/s$, $\hat{S}p = r_0/r$, $\hat{p} = s/n$, $\text{Var}(\hat{S}e) = \hat{S}e(1 - \hat{S}e)/s$, $\text{Var}(\hat{S}p) = \hat{S}p(1 - \hat{S}p)/r$, $\text{Var}(\hat{p}) = \hat{p}(1 - \hat{p})/n$, $\hat{Y} = (s_1r + r_0s - sr)/sr$ y $\hat{Q} = (s_1 + r_1)/n$.

Cuando $s_0 = r_1$ las varianzas asintóticas estimadas son

$$\bar{\text{Var}}(\hat{\kappa}_1) = \bar{\text{Var}}(\hat{\kappa}_2) = \bar{\text{Var}}(\hat{Y}) = \frac{\hat{S}e(1 - \hat{S}e)}{s} + \frac{\hat{S}p(1 - \hat{S}p)}{r}.$$

Para los dos coeficientes kappa promedio se han estudiado los siguientes intervalos de confianza asintóticos (similares a los del coeficiente kappa ponderado).

a). Intervalo de confianza tipo Wald

Asumiendo la normalidad asintótica de $\hat{\kappa}_1$ and $\hat{\kappa}_2$, el intervalo de confianza cada coeficiente kappa promedio es

$$\hat{\kappa} \pm z_{1-\alpha/2} \sqrt{\text{Var}(\hat{\kappa})},$$

donde $\hat{\kappa}$ es $\hat{\kappa}_1$ o $\hat{\kappa}_2$.

b). Intervalo de confianza bootstrap

La estimación de los coeficientes kappa promedio se puede realizar aplicando el método bootstrap, de forma similar a como se ha explicado para el coeficiente kappa ponderado. A partir de la muestra aleatoria observada se generan K muestras con reemplazamiento, y a partir de estas K muestras se calculan los intervalos de confianza. Por consiguiente, como estimador de cada coeficiente kappa promedio se propone el valor promedio obtenido con las K muestras aleatorias, y a continuación se calculan a partir de las K muestras el intervalo de confianza corregido por el sesgo [8] para cada uno de los coeficientes kappa promedio.

c). Intervalo de confianza logit

Como la evaluación del coeficiente kappa ponderado de un test diagnóstico se limita a los valores positivos (entre 0 and 1) de este parámetro, los valores de los coeficientes kappa promedio también están entre 0 y 1, pudiéndose realizar la transformación logit de estos parámetros. Así, asumiendo la normalidad asintótica de $\hat{\kappa}$, la transformación logit de $\hat{\kappa}$, $\ln[\hat{\kappa}/(1-\hat{\kappa})]$, sigue una distribución normal de media $\ln[\kappa/(1-\kappa)]$. De esta forma intervalo de confianza para el logit de κ es

$$\text{logit}(\hat{\kappa}) \pm z_{1-\alpha/2} \sqrt{\text{Var}(\text{logit}(\hat{\kappa}))},$$

siendo las varianzas asintóticas estimadas cuando $s_0 \neq r_1$

$$\begin{aligned} \text{Var}(\text{logit}(\hat{\kappa}_i)) = & \frac{1}{(\hat{\kappa}(0) + \hat{\kappa}(1))^2 [\hat{\kappa}(0) - \hat{\kappa}(1)]^2 \hat{\kappa}_i^2 [1 - \hat{\kappa}_i^2]^2} \times \\ & \left\{ \frac{\left[\hat{\kappa}(1) [2(1 - \hat{\kappa}_i) \hat{\kappa}(0)^2 - (\hat{\kappa}(0)(1 - 2\hat{\kappa}(0)) + \hat{\kappa}(1)) \hat{\kappa}_i] \right]^2 (1 - \hat{S}p)^2 \hat{Y}^2 \text{Var}(\hat{p}) + \hat{p}^2 [(1 - \hat{S}p) \text{Var}(\hat{S}e) + \hat{S}e^2 \text{Var}(\hat{S}p)]}{\hat{\kappa}(0)} \right\} \frac{1}{\hat{Q}^4} + \\ & \left\{ \frac{\left[\hat{\kappa}(0) [(\hat{\kappa}(0) + \hat{\kappa}(1)) \hat{\kappa}_i - 2\hat{\kappa}(0) \hat{\kappa}(1)] \right]^2 (1 - \hat{S}e)^2 \hat{Y}^2 \text{Var}(\hat{p}) + (1 - \hat{p})^2 [\hat{S}p \text{Var}(\hat{S}e) + (1 - \hat{S}e)^2 \text{Var}(\hat{S}p)]}{\hat{\kappa}(1)} \right\} \frac{1}{(1 - \hat{Q})^4} + \\ & 2 \left\{ \frac{\left[\hat{\kappa}(1) [2(1 - \hat{\kappa}_i) \hat{\kappa}(0)^2 - (\hat{\kappa}(0)(1 - 2\hat{\kappa}(0)) + \hat{\kappa}(1)) \hat{\kappa}_i] \right]}{\hat{\kappa}(0)} \right\} \left\{ \frac{\left[\hat{\kappa}(0) [(\hat{\kappa}(0) + \hat{\kappa}(1)) \hat{\kappa}_i - 2\hat{\kappa}(0) \hat{\kappa}(1)] \right]}{\hat{\kappa}(1)} \right\} \times \\ & \left. \frac{(1 - \hat{p}) \hat{p} [(1 - \hat{S}e) \hat{S}e \text{Var}(\hat{S}p) + (1 - \hat{S}p) \hat{S}p \text{Var}(\hat{S}e)] - (1 - \hat{S}e)(1 - \hat{S}p) \hat{Y}^2 \text{Var}(\hat{p})}{\hat{Q}^2 (1 - \hat{Q})^2} \right\}. \end{aligned}$$

y

$$\begin{aligned} \bar{Var}(\text{logit}(\hat{\kappa}_2)) &= \frac{1}{[\hat{\kappa}(0) + \hat{\kappa}(1)]^2 [\hat{\kappa}(0) - \hat{\kappa}(1)]^2 \hat{\kappa}_2^2 [1 - \hat{\kappa}_2^2]^2} \times \\ &\left\{ \frac{\hat{\kappa}(1) [2\hat{\kappa}(0)\hat{\kappa}(1) - (\hat{\kappa}(0) + \hat{\kappa}(1))\hat{\kappa}_2]}{\hat{\kappa}(0)} \right\}^2 \frac{(1 - \hat{S}p)^2 \hat{Y}^2 \bar{Var}(\hat{p}) + \hat{p}^2 [(1 - \hat{S}p)\bar{Var}(\hat{S}e) + \hat{S}e^2 \bar{Var}(\hat{S}p)]}{\hat{Q}^4} + \\ &\left\{ \frac{\hat{\kappa}(0) [(\hat{\kappa}(0) + \hat{\kappa}(1))\hat{\kappa}_2 - 2\hat{\kappa}(1)^2]}{\hat{\kappa}(1)} \right\}^2 \frac{(1 - \hat{S}e)^2 \hat{Y}^2 \bar{Var}(\hat{p}) + (1 - \hat{p})^2 [\hat{S}p\bar{Var}(\hat{S}e) + (1 - \hat{S}e)^2 \bar{Var}(\hat{S}p)]}{(1 - \hat{Q})^4} + \\ &2 \left\{ \frac{\hat{\kappa}(1) [2\hat{\kappa}(0)\hat{\kappa}(1) - (\hat{\kappa}(0) + \hat{\kappa}(1))\hat{\kappa}_2]}{\hat{\kappa}(0)} \right\} \left\{ \frac{\hat{\kappa}(0) [(\hat{\kappa}(0) + \hat{\kappa}(1))\hat{\kappa}_2 - 2\hat{\kappa}(1)^2]}{\hat{\kappa}(1)} \right\} \times \\ &\frac{(1 - \hat{p})\hat{p} [(1 - \hat{S}e)\hat{S}e\bar{Var}(\hat{S}p) + (1 - \hat{S}p)\hat{S}p\bar{Var}(\hat{S}e)] - (1 - \hat{S}e)(1 - \hat{S}p)\hat{Y}^2 \bar{Var}(\hat{p})}{\hat{Q}^2 (1 - \hat{Q})^2}, \end{aligned}$$

respectivamente, y cuando $s_0 = r_1$ las varianzas asintóticas estimadas son

$$\bar{Var}(\log(\hat{\kappa}_1)) = \bar{Var}(\log(\hat{\kappa}_2)) = \bar{Var}(\log(\hat{Y})) = \frac{1}{\hat{Y}(1 - \hat{Y})} \left(\frac{\hat{S}e(1 - \hat{S}e)}{s} + \frac{\hat{S}p(1 - \hat{S}p)}{r} \right).$$

Finalmente, el intervalo de confianza logit para cada coeficiente kappa promedio es

$$\left(\frac{\exp\left\{\text{logit}(\hat{\kappa}) - z_{1-\alpha/2} \sqrt{\bar{Var}(\text{logit}(\hat{\kappa}))}\right\}}{1 + \exp\left\{\text{logit}(\hat{\kappa}) - z_{1-\alpha/2} \sqrt{\bar{Var}(\text{logit}(\hat{\kappa}))}\right\}}, \frac{\exp\left\{\text{logit}(\hat{\kappa}) + z_{1-\alpha/2} \sqrt{\bar{Var}(\text{logit}(\hat{\kappa}))}\right\}}{1 + \exp\left\{\text{logit}(\hat{\kappa}) + z_{1-\alpha/2} \sqrt{\bar{Var}(\text{logit}(\hat{\kappa}))}\right\}} \right)$$

donde $\hat{\kappa}$ es $\hat{\kappa}_1$ o $\hat{\kappa}_2$.

Como criterios de aplicación se puede establecer, en términos generales, la siguiente regla:

- Para tamaños muestrales de entre 100 y 500 individuos, aplicar el intervalo tipo Wald.
- Para tamaños muestrales superiores a 500, se puede aplicar cualquiera de los tres intervalos, si bien el intervalo bootstrap requiere un mayor esfuerzo computacional.

3. EJEMPLO

Los resultados obtenidos se han aplicado al estudio de Weiner et al [9] sobre el diagnóstico de la enfermedad de la arteria coronaria, utilizando como test diagnóstico un test de ejercicio y como gold estándar la arteriografía coronaria. En la Tabla2 se muestran los resultados obtenidos por Weiner et al para individuos con angina de pecho, y donde la variable T modeliza el resultado del test de ejercicio y la variable D el resultado de la angiografía coronaria.

Tabla 2. Datos del estudio de Weiner et al (1979).

Frecuencias observadas			
	<i>T</i> = 1	<i>T</i> = 0	Total
<i>D</i> = 1	473	81	554
<i>D</i> = 0	22	44	66
Total	495	125	620

En la Tabla 3 se muestran los valores estimados de cada parámetro y los intervalos de confianza al 95%. La interpretación de los resultados (en términos de las estimaciones puntuales) es la siguiente:

a). Sensibilidad y especificidad. La sensibilidad estimada toma un valor moderadamente alto, por lo que el test diagnóstico (test de ejercicio) es útil para descartar la enfermedad. La especificidad estimada toma un valor moderado, por lo que el test de ejercicio es moderadamente útil para confirmar la enfermedad en individuos que son sospechosos de tenerla.

b). Razones de verosimilitud. La probabilidad de que el test de ejercicio sea positivo es 2.561 veces mayor en los individuos con la enfermedad de la arteria coronaria que en los individuos sin esta enfermedad. La probabilidad de que el test de ejercicio sea negativo es 4.566 ($=1/0.219$) veces mayor en los individuos sin la enfermedad de la arteria coronaria que en los individuos con esta enfermedad.

c). Valores predictivos. El valor predictivo positivo estimado toma un valor muy alto, por lo que el test de ejercicio es muy bueno para confirmar la enfermedad de la arteria coronaria en los individuos pertenecientes a la población objeto de estudio. El valor predictivo negativo estimado toma un valor bajo, por lo que el test de ejercicio no es un test adecuado para descartar la enfermedad en los individuos pertenecientes a la población objeto de estudio.

Tabla 3. Resultados del estudio de Weiner et al (1979).

Sensibilidad y especificidad		
	Estimación	IC al 95%
<i>Se</i>	0.854	(0.822 , 0.881)
<i>Sp</i>	0.667	(0.547 , 0.769)
Razones de verosimilitud		
	Estimación	IC al 95%
<i>LR</i> ⁺	2.561	(1.871 , 3.665)
<i>LR</i> ⁻	0.219	(0.172 , 0.290)
Valores predictivos		
	Estimación	IC al 95%
<i>VPP</i>	0.956	(0.934 , 0.971)
<i>VPN</i>	0.352	(0.274 , 0.439)
Coeficiente kappa promedio		
	Estimación	IC logit al 95%
κ_1	0.463	(0.360 , 0.568)
κ_2	0.319	(0.239 , 0.412)
Coeficiente kappa ponderado		

c	$\hat{\kappa}(c)$	IC logit al 95%
0.1	0.527	(0.408 , 0.639)
0.2	0.462	(0.370 , 0.584)
0.3	0.412	(0.338 , 0.539)
0.4	0.371	(0.310 , 0.502)
0.5	0.338	(0.285 , 0.471)
0.6	0.310	(0.264 , 0.444)
0.7	0.287	(0.245 , 0.420)
0.8	0.267	(0.229 , 0.399)
0.9	0.249	(0.214 , 0.380)
Coeficiente kappa promedio		
	Estimación	IC logit al 95%
κ_1	0.463	(0.360 , 0.568)
κ_2	0.319	(0.239 , 0.412)

d). Coeficiente kappa ponderado. Si el test de ejercicio se utiliza como test previo a un tratamiento de riesgo ($0 < c < 0.5$), el coeficiente kappa ponderado estimado toma un valor intermedio para cada valor de c y el acuerdo más allá del azar entre el test de ejercicio y la angiografía coronaria (gold estándar) es principalmente moderado. Si el test de ejercicio se utiliza para un diagnóstico simple ($c = 0.5$), el acuerdo más allá del azar entre el test de ejercicio y la angiografía coronaria varía entre mediocre y moderado; y si el test de ejercicio se utiliza como un test de screening ($0.5 < c < 1$), el acuerdo más allá del azar entre el test de ejercicio y la angiografía coronaria es principalmente mediocre.

e). Coeficiente kappa promedio. Si el clínico considera que la pérdida asociada a un falso positivo (L') es mayor que la pérdida asociada a un falso negativo (L), el valor estimado del coeficiente kappa promedio es moderado, y por tanto el acuerdo promedio más allá del azar entre el test de ejercicio y la angiografía es moderada en la población objeto de estudio. Si el clínico considera que la pérdida asociada a un falso negativo (L) es mayor que la pérdida asociada a un falso positivo (L'), el valor estimado del coeficiente kappa promedio es mediocre, y por tanto el acuerdo promedio más allá del azar entre el test de ejercicio y la angiografía es mediocre en la población objeto de estudio. Por tanto, considerando las pérdidas, el test de ejercicio no debería de aplicarse como test de screening en la población objeto de estudio, y si se puede aplicar como test previo a un tratamiento de riesgo en la población objeto de estudio.

4. CONCLUSIONES

La utilización de los tests diagnósticos se ha convertido en una herramienta fundamental en la práctica clínica, por lo que la Estadística ha ido desarrollando métodos para ir resolviendo los problemas que la aplicación práctica de los tests diagnósticos ha dado lugar. En este trabajo se ha realizado una revisión de los parámetros más usuales (sensibilidad, especificidad, razones de verosimilitud y valores predictivos) y de otros más novedosos (coeficiente kappa ponderado y coeficiente kappa promedio) de un test diagnóstico binario, exponiéndose los mejores intervalos de confianza para cada uno de ellos bajo un muestreo transversal. Todos estos parámetros no son alternativos unos a otros, sino que son parámetros complementarios que, conjuntamente, ayudan a entender el mecanismo de clasificación de un test diagnóstico binario. El presente trabajo tiene como principal objetivo orientar al investigador aplicado (clínico, epidemiólogo,...) sobre la interpretación y aplicación práctica de estos parámetros, y también a aquellos otros investigadores (estadísticos o matemáticos) que deseen introducirse en este ámbito de la Bioestadística.

AGRADECIMIENTOS

Este trabajo ha sido financiado por la Subdirección General de Proyectos de Investigación del Ministerio de Economía y Competitividad, España, Proyecto MTM2012-35591. También quiero dar las gracias al profesor Carlos Bouza y al resto de editores, sin cuyo esfuerzo este libro no sería posible.

REFERENCIAS

1. ZHOU, X.H., Obuchowski, N., McClish, D.,(2002):Statistical Methods in Diagnostic Medicine. John Wiley and Sons: New York.
2. PEPE, M.S.,(2003):The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press: New York.
3. KRAEMER, H.C., (1992): Evaluating medical tests. SAGE Publications, Newbury Park.
4. ROLDÁN-NOFUENTES, J.A., OLVERA-PORCEL, C., (2014): Average kappa coefficient: a new measure to assess a binary test considering the losses associated with an erroneous classification. **Journal of Statistical Computation and Simulation**, in press, doi: 10.1080/00949655.2014.881816.
5. YU, W., GOU, X., XU, W., (2014): An improved score interval with a modified midpoint for a binomial proportion. **Journal of Statistical Computation and Simulation** 84, 1022-1038.
6. MARTÍN-ANDRÉS, A., ÁLVAREZ-HERNÁNDEZ, M., (2014): Two-tailed approximate confidence intervals for the ratio of proportions. **Statistics and Computing**,24, 65-75.
7. ROLDÁN NOFUENTES, J.A., LUNA DEL CASTILLO, J.D. and MONTERO ALONSO, M.A., (2009): Confidence intervals of weighted kappa coefficient of a binary diagnostic test. **Communications in Statistics - Simulation and Computation** 38, 1562-1578.
8. EFRON, B., TIBSHIRANI, R.J., (1993): An Introduction to the Bootstrap. Chapman and Hall: New York.
9. WEINER, D. A., RYAN, T. J., MCCABE, C. H., KENNEDY, J. W., SCHLOSS, M., TRISTANI, F., CHAITMAN, B. R., FISHER, L. D., (1979): Exercise stress testing. Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the coronary artery surgery study (CASS). **The New England Journal of Medicine** 301, 230-235.

Capítulo 9

CLASIFICACIÓN USANDO REDES NEURONALES ARTIFICIALES

Luis René Marcial Castillo, Marcela Rivera Martínez, Lourdes Sandoval Solís
Benemérita Universidad Autónoma de Puebla
Puebla, Puebla, México

lmc@cs.buap.mx, mrivmar@hotmail.com, sandoval@cs.buap.mx

Resumen: *La clasificación es una tarea muy importante en diversas aplicaciones de la salud y del medio ambiente. En este trabajo, se realiza la clasificación mediante el uso de tres modelos de redes neuronales artificiales: red neuronal multicapa, red neuronal en base radial y red neuronal probabilística. Se desarrollan las pruebas usando las bases de datos de uso libre: fertilidad, parkinson, cambio climático y plantas.*

Palabras Clave: Clasificación, red neuronal multicapa, red neuronal en base radial, red neuronal probabilística, aprendizaje supervisado.

1. Introducción

El problema de clasificación continúa siendo un problema muy importante como lo muestran varios trabajos recientes. En el área de la salud, para clasificar Arritmias cardiacas, que son alteraciones del ritmo del corazón, caracterizadas por una irregularidad de frecuencia y una desigualdad de sus contracciones (Khoshnoud y Ebrahimnezhad2013), para clasificar señales epilépticas que surgen a partir de un electro encefalograma conocidas también como señales de tipo EEG (Goshvarpour1, Ebrahimnezhad, y Goshvarpour,2013), para clasificar la calidad del líquido seminal (Gila, Girelab, De Juan, Gomez y Johnssonc, 2012) o para clasificar pacientes con la enfermedad de Parkinson tomando en cuenta parámetros sobre la calidad del habla (Erdogdu, 2013). En el área del medio ambiente, para clasificar fallos en los modelos climáticos que se han usado en la predicción (Lucas, Klein, Tannahill, Ivanova, Brandon, Domyancic y Zhang, 2013) o para clasificar diferentes tipos de plantas (Mallah, Cope y Orwell, 2013). El presente trabajo, explica los modelos de redes neuronales y aplica las librerías de OCTAVE para resolver el problema de clasificación mediante redes neuronales con aprendizaje supervisado, las pruebas se realizan con problemas prueba (benchmark) de bases de datos de uso libre.

El trabajo está organizado de la manera siguiente. En la sección 2, se presentan y explican los modelos de redes neuronales artificiales: red neuronal multicapa, red neuronal en base radial y red neuronal probabilística, en la sección 3, se presentan los experimentos y resultados usando para ello cuatro problemas prueba, la sección 4 presenta las conclusiones y finalmente se listan las referencias.

2. Redes neuronales artificiales

Las redes neuronales artificiales (RNA) intentan modelar las capacidades de procesamiento del sistema nervioso, por lo que se deben tener en cuenta las propiedades esenciales de las redes neuronales biológicas desde la perspectiva de su procesamiento. Esta información permite diseñar modelos abstractos de redes neuronales artificiales que pueden ser simulados y analizados (Rojas, 2007). Las RNA presentan una serie de características propias del cerebro: aprenden mediante el entrenamiento y resuelven problemas similares con los que nunca se han entrenado. Se puede definir formalmente a una red neuronal artificial como:

Definición 1. Una red neuronal es una tripleta (N, V, w) con dos conjuntos N, V y una función w , donde N es el conjunto de neuronas y V es un conjunto $\{(i, j) \mid i, j \in Z^+\}$ cuyos elementos son denominadas conexiones entre la neurona i y la neurona j . La función $w: V \rightarrow \mathcal{R}$ define los pesos, donde $w(i, j)$ es el peso entre la neurona i y la neurona j .

Cualquier modelo de red neuronal artificial contiene tres elementos esenciales (Rojas, 1996):

1. La **estructura de sus neuronas**, es decir, el contenido de la neurona.
2. La **topología de la red neuronal**, es decir, la forma en la que se realizan las conexiones de las neuronas.
3. El **algoritmo de aprendizaje**, es decir, el método utilizado para encontrar los valores óptimos de los pesos que se involucran en la red neuronal artificial.

A continuación se explican los modelos de redes neuronales artificiales usados en la clasificación de las bases de datos.

2.1 Red neuronal multicapa

Las redes multicapa son aquellas en las cuales el conjunto de neuronas N es subdividido en m subconjuntos N_1, N_2, \dots, N_m de tal manera que forman un grupo de capas simples en cascada, es decir, las salidas de la capa N_1 son conectadas como entrada a la capa N_2 , las salidas de la capa N_2 son las entradas de la capa N_3 , y así sucesivamente. Las entradas x_1, \dots, x_e del problema son conectadas a la capa N_1 y las salidas del problema t_1, \dots, t_s son las salidas de la capa N_m . El esquema de la red neuronal multicapa se muestra en la figura 1. No existe un resultado que nos diga para un problema específico cuántas capas ocultas colocar ni cuántas neuronas agregar por capa oculta, la cantidad se especifica a prueba y error. Los pesos o parámetros que deberán ajustarse para obtener la salida deseada en el aprendizaje supervisado se representan por las matrices W_1, W_2, \dots, W_m . Por ejemplo, supóngase que se desea clasificar rostros humanos en base a su sexo (masculino o femenino), si la imagen digital que representa el rostro humano es de $64 \times 64 = 4096$ píxeles en tonos de gris, entonces la capa de entrada contendrá 4096 sitios de entrada y la capa de salida contendrá 2 neuronas si la salida para femenino se representa con $[1 \ 0]$ y para masculino con $[0 \ 1]$, si además se supone que existen 3 capas ocultas una con 20 neuronas, otra con 10 y la última con 5 neuronas, entonces W_1 contendrá $4096 \times 20 = 81920$ pesos, W_2 contendrá $20 \times 10 = 200$ pesos, W_3 contendrá $10 \times 5 = 50$ pesos y W_4 contendrá $5 \times 2 = 10$ pesos. Por lo tanto, la cantidad de pesos o parámetros que deberán ajustarse es $81920 + 200 + 50 + 10 = 82180$, es decir, se tendría que resolver un problema de optimización de 82180 variables.

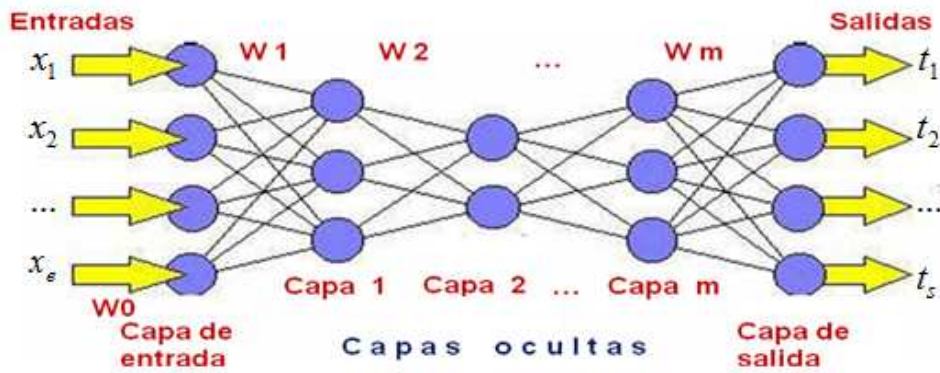


Figura 1. Red neuronal multicapa.

Si al construir la función que mide el error se tiene que ésta es continuamente diferenciable, entonces, se puede usar algún método de búsqueda lineal para encontrar los pesos óptimos. En caso de que la función de error no sea diferenciable se puede utilizar alguna meta heurística para ajustar los pesos.

Para encontrar los pesos óptimos de una red neuronal multicapa Rumelhart, Hilton y Williams propusieron en 1986 el algoritmo conocido como retro propagación hacia atrás (backpropagation). El algoritmo de aprendizaje backpropagation conlleva una fase de propagación hacia adelante y otra fase de propagación hacia atrás. El algoritmo es el siguiente (Rojas, 1996):

Considérese una red neuronal artificial con una sola entrada x y una función de red $f(x)$, la derivada $f'(x)$ se calcula en dos fases:

Propagación hacia adelante. En esta fase, la entrada x es alimentada a la red, se evalúan y almacenan los valores obtenidos de las funciones primitivas y sus derivadas en cada uno de los nodos, el valor de la función se almacena en la parte derecha del nodo y el valor de la derivada se almacena en la parte izquierda.

Propagación hacia atrás. La constante uno se alimenta en la unidad de salida y la red es alimentada en orden inverso. La información en un nodo se suma y el resultado se multiplica por el valor almacenado en la parte izquierda de cada unidad. El resultado se transmite a la izquierda de la unidad. Finalmente, el resultado coleccionado en la unidad de entrada, es la derivada de la función de la red neuronal con respecto a la variable x .

Ejemplo 1. Supóngase que se desea clasificar la función booleana de 2 variables $f(x,y)$ conocida como la función *xor*, ilustrada en el cuadro 1.

Cuadro1. Función booleana *xor* de dos variables.

x	y	$f(x,y)=xor(x,y)$
0	0	0
0	1	1
1	0	1
1	1	0

La arquitectura usada es la que se muestra en la figura 2.

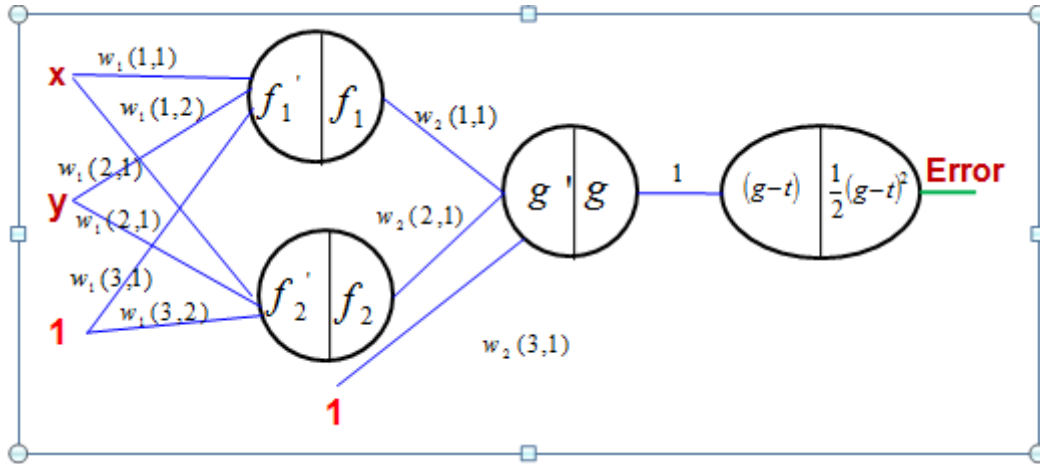


Figura 2. Red neuronal multicapa para el problema *xor*.

Para este problema deberán ajustarse 9 parámetros. Si se usa como función de activación a la función sigmoide dada por:

$$f(x) = \frac{1}{1 + e^{-x}}. \quad (1)$$

Cuya derivada se expresa en términos de sí misma, es decir,

$$f'(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{1 + e^{-x}} \left(1 - \frac{1}{1 + e^{-x}} \right) = f(x)(1 - f(x)). \quad (2)$$

Las derivadas parciales con respecto a la función de error se evalúan como:

$$\frac{\partial \text{Error}(W1, W2)}{\partial w_2(1,1)} = (g-t)g'f_1. \quad (3)$$

$$\frac{\partial \text{Error}(W1, W2)}{\partial w_2(2,1)} = (g-t)g'f_2. \quad (4)$$

$$\frac{\partial \text{Error}(W1, W2)}{\partial w_2(3,1)} = (g-t)g'f_3. \quad (5)$$

$$\frac{\partial \text{Error}(W1, W2)}{\partial w_1(1,1)} = (g-t)g'w_2(1,1)f_1'x. \quad (6)$$

$$\frac{\partial \text{Error}(W1, W2)}{\partial w_1(1,2)} = (g-t)g'w_2(2,1)f_2'x. \quad (7)$$

$$\frac{\partial \text{Error}(W1, W2)}{\partial w_1(2,1)} = (g-t)g'w_2(1,1)f_1'y. \quad (8)$$

$$\frac{\partial \text{Error}(W1, W2)}{\partial w_1(2,2)} = (g-t)g'w_2(2,1)f_2'y. \quad (9)$$

$$\frac{\partial \text{Error}(W1, W2)}{\partial w_1(3,1)} = (g-t)g'w_2(1,1)f_1' * 1. \quad (10)$$

$$\frac{\partial \text{Error}(W1, W2)}{\partial w_1(3,2)} = (g-t)g'w_2(2,1)f_2' * 1. \quad (11)$$

Para encontrar los pesos óptimos de los parámetros almacenados en forma matricial

por $W_1 = \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \\ w_{3,1} & w_{3,2} \end{bmatrix}$ y $W_2 = \begin{bmatrix} w_{1,1} \\ w_{2,1} \\ w_{3,1} \end{bmatrix}$, mediante el método de máximo descenso se

siguen los pasos siguientes:

1. Generar una aproximación inicial de forma aleatoria para las marices W^0_1 y W^0_2 .
2. Calcular el gradiente $grad_0$ de W^0_1 y W^0_2 usando las ecuaciones (3)-(11) e inicializar k con el valor de 1.
3. Poner la dirección de descenso $d_k = -grad_k$.
4. Mientras $grad_k \neq 0$
 - a. Calcular el tamaño de paso α_k mediante una búsqueda lineal inexacta o bien asignarle un valor suficientemente pequeño.
 - b. Calcular $W^{k+1}_1 = W^k_1 + \alpha_k * d_k$ y $W^{k+1}_2 = W^k_2 + \alpha_k * d_k$.
 - c. Calcular el gradiente $grad_{k+1}$ usando las ecuaciones (3)- (11).
 - d. Poner $d_{k+1} = -grad_{k+1}$ y $k = k + 1$.

La ejecución de la implementación computacional del algoritmo que no usa librería alguna para redes neuronales y realizada con la computadora especificada en la sección 3, entregó los resultados siguientes para los pesos óptimos:

$$W_1 = \begin{bmatrix} -5.7338 & -3.6772 \\ -5.6286 & -3.6610 \\ 2.0758 & 5.3744 \end{bmatrix}, \quad W_2 = \begin{bmatrix} -7.5248 \\ 7.1598 \\ -3.2273 \end{bmatrix}.$$

Tales valores de los pesos corresponden al esquema que se muestra en la figura 3, y que al evaluar la red neuronal se obtiene la función booleana *xor* en dos variables.

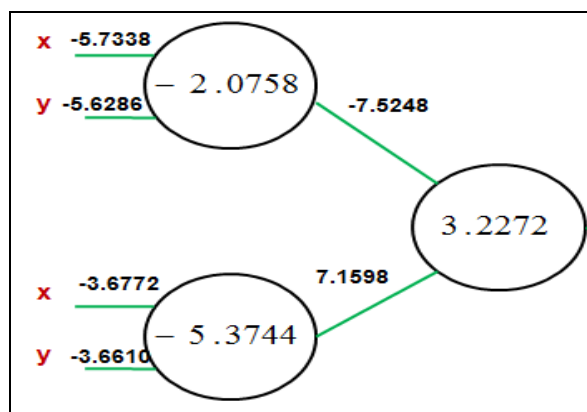


Figura 3. Solución de la función booleana *xor* de dos variables.

La dirección de descenso también puede ser la dirección de gradientes conjugados para problemas de alta dimensión y la dirección de newton o cuasi newton si el problema es de baja dimensión.

2.2 Red neuronal en base radial

Las redes neuronales en funciones de base radial (FBR) surgen de la teoría de aproximación de funciones y fueron usadas para el reconocimiento de patrones (Tou y

González, 1974). Estas redes neuronales constan de dos capas, donde los nodos ocultos implementan un conjunto de funciones en base radial y los nodos de salida implementan funciones de sumatoria lineales, es decir, la entrada en estas redes es no lineal y la salida es lineal. Debido a la característica no lineal, las redes neuronales en base radial son hábiles para modelar mapeos complejos, cosa que el perceptrón multicapa puede solo realizar mediante múltiples capas intermedias. El entrenamiento para estas redes es dividido en dos estados: primero se determinan los pesos que conectan las entradas con la capa oculta y después se determinan los pesos de la capa oculta a la capa de salida, el entrenamiento en estas redes es muy rápido.

Ejemplo 2. Supóngase que se desea clasificar mediante una red neuronal en base radial a la función booleana de 2 variables $f(x,y)$ conocida como la función *xor*, ilustrada en el cuadro 1. La estructura de la red neuronal en base radial para este problema se ilustra en la figura 4. La función de activación en la capa oculta es la función Gaussiana dada por la expresión

$$f(x) = e^{-\frac{x^2}{2\sigma^2}}. \quad (12)$$

En la capa de salida se usa la función lineal dada por la expresión

$$g(x) = x. \quad (13)$$

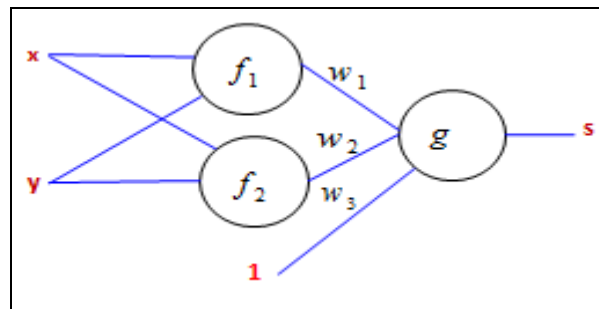


Figura 4. Topología usada de la red neuronal en base radial para el *xor* de dos variables.

Para encontrar los pesos óptimos en la red neuronal en base radial para la arquitectura de la figura 4, se deben seguir los pasos siguientes:

Paso 1. Encontrar los centros de los datos. Dos centros para el caso de 2 neuronas en la capa oculta, la forma más sencilla es utilizar el algoritmo k medias.

Paso 2. Para cada nodo de la capa oculta encontrar la distancia euclidiana entre los datos y el centro asignado al nodo.

Paso 3. Para cada nodo de la capa oculta aplicar la función Gaussiana.

Paso 4. Encontrar w resolviendo el sistema lineal $Gw=s$, donde s es el vector de salidas que dicta el entrenamiento y G es la matriz que se debe obtener en el paso 3.

En caso de no encontrar la solución del problema se deberá incrementar la cantidad de nodos en la capa oculta.

2.3 Red neuronal probabilística

Una red neuronal probabilista es la implementación de un algoritmo estadístico denominado análisis de discriminante en el cual las operaciones son organizadas en una red con cuatro capas (Chen y Tsai, 2014): capa de entrada, capa de patrones, capa de

sumatoria y capa de salida. La arquitectura básica de una red neuronal probabilística se ilustra en la figura 5. La capa de entrada contendrá los valores de los atributos que se presentan a la red en cada entrada, la capa de patrones calculará la distancia entre el vector de entrada y la salida esperada produciendo un vector cuyos elementos mostrarán la cercanía entre la entrada y la salida, la capa de sumatoria sumará las contribuciones para cada clase de entradas con el fin de producir un vector de salida que contendrá las probabilidades y finalmente la capa de salida contendrá una cantidad de neuronas igual a la cantidad de clases que se desea clasificar.

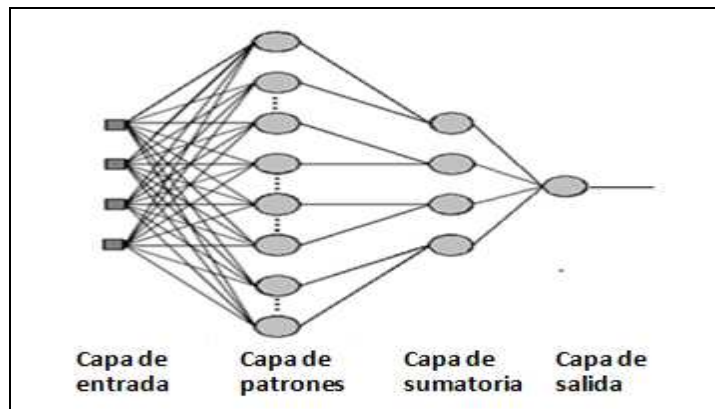


Figura 5. Topología de una red neuronal probabilística.

3. Experimentos y resultados

Se realizó la clasificación de cuatro problemas prueba dos del área de la salud y dos del medio ambiente. Para la clasificación se utilizaron las librerías de OCTAVE probando con redes multicapa, redes en base radial y redes probabilísticas. Todas las corridas se realizaron en una computadora Intel Core i7 con 2.20 GHz y 8.0GB de memoria RAM.

El nombre de los problemas prueba, así como la cantidad de instancias, atributos y clases de cada uno de ellos se muestran en el cuadro 2.

Cuadro 2. Problemas benchmark.

Nombre	Instancias	Atributos	Clases
Fertilidad	100	10	2
Parkinson	1040	26	2
Modelo climático	540	18	2
Plantas	1600	64	100

El problema prueba llamado *Fertilidad* (Gila, Girelab, De Juan, Gómez y Johnssonc, 2012), se diseñó en el año 2010 (Gil y Girela, 2013). Esta base de datos ofrece los resultados de muestras de líquido seminal obtenidos de 100 voluntarios que fueron analizadas bajo la norma del año 2010 de la Organización Mundial de Salud, los atributos son asociados a datos socio-demográficos, factores ambientales, estado de salud y hábitos de vida. El problema prueba *Parkinson* (Erdogdu, Isenkul, Sakar, Sertbas, Gurgen, Delil, Apaydin y Kursun, 2013), se diseñó en el año 2013. Esta base de datos consta de datos de entrenamiento de individuos que tienen o no el mal de Parkinson, en donde se toman múltiples tipos de grabaciones de sonido incluyendo las vocales, números, palabras y oraciones cortas. El problema prueba *Modelo climático*

(Lucas, Klein, Tannahill, Ivanova, Brandon, Domyancic y Zhang2013) apareció en el año 2013 (Lucas, 2013); los atributos son asociados a accidentes de simulación de los modelos climáticos cuando se usa el modelo estadístico hiper cubo latino. Finalmente, el problema prueba *Plantas* (Mallah, Cope y Orwell, 2013; Mallah y Orwell 2013) apareció en el año 2012 (Mallah, C. 2012), en esta base de datos los atributos están asociados a la forma, textura o margen de una planta.

Para el caso de la red neuronal multicapa se decidió a prueba y error, la cantidad de capas ocultas a utilizar y cuántas neuronas ubicar en cada capa oculta. Se observó que la arquitectura que daba un menor error cuadrático medio (MSE) fue cuando se tenían tres capas ocultas en todos los casos. Para las bases de datos *Fertilidad* y *Modelo climático*, la primera capa oculta contiene 10 neuronas, la segunda y tercera contienen 5 neuronas. Para la base de datos *Plantas* y *Parkinson*, las tres capas ocultas contienen 64 neuronas.

Todos los modelos de redes neuronales se entrenaron con el 50% de los datos y se probó su predicción en el 50% restante.

Para la base de datos *Fertilidad*, el cuadro 3 muestra los resultados del entrenamiento para cada tipo de red neuronal: cantidad de neuronas utilizadas, tiempo de entrenamiento, porcentaje de aciertos y error cuadrático medio (MSE). En el cuadro 4, se muestran los resultados de la predicción: porcentaje de aciertos, tiempo de ejecución, así como el MSE.

Cuadro 3. Resultados del entrenamiento para el problema prueba *Fertilidad*.

Tipo de red neuronal	Neuronas	Tiempo	Porcentaje de aciertos	MSE
Multicapa	21	22.690seg.	94	0.0507
Radial	80	0.236seg.	100	0.0
Probabilística	11	0.040seg.	100	0.0

Cuadro 4. Resultados de la predicción para el problema prueba *Fertilidad*.

Tipo de red neuronal	Tiempo	Porcentaje de aciertos	MSE
Multicapa	0.011seg.	84	0.16
Radial	0.103seg.	80	0.2
Probabilística	0.008seg.	78	0.22

Para la base de datos *Parkinson*, el cuadro 5 muestra los resultados del entrenamiento para cada tipo de red neuronal: cantidad de neuronas utilizadas, tiempo de entrenamiento, porcentaje de aciertos y error cuadrático medio (MSE) y el cuadro 6, muestra los resultados de la predicción.

Cuadro 5. Resultados del entrenamiento para el problema prueba *Parkinson*.

Tipo de red neuronal	Neuronas	Tiempo	Porcentaje de aciertos	MSE
Multicapa	21	6.30seg.	100	0.00000
Radial	500	36.34seg.	98.0769	0.00005
Probabilística	27	0.082seg.	100	0.00000

Cuadro 6. Resultados de la predicción para el problema prueba *Parkinson*.

Tipo de red neuronal	Tiempo	Porcentaje de aciertos	MSE
Multicapa	0.009seg.	100	0.0000
Radial	0.04seg.	68.07	0.28
Probabilística	0.07seg.	70.1923	0.2990

Para la base de datos *Modelo climático*, el cuadro 7 muestra los resultados del entrenamiento para cada tipo de red neuronal: cantidad de neuronas utilizadas, tiempo de entrenamiento, porcentaje de aciertos y error cuadrático medio (MSE). En el cuadro 8, se muestran los resultados de la predicción: porcentaje de aciertos, tiempo de ejecución, así como el MSE.

Cuadro 7. Resultados del entrenamiento para el problema prueba *Modelo climático*.

Tipo de red neuronal	Neuronas	Tiempo	Porcentaje de aciertos	MSE
Multicapa	21	4.7222seg.	96.66	0.0261
Radial	80	0.0753seg.	100	0.0
Probabilística	19	0.0479seg.	100	0.0

Cuadro 8. Resultados de la predicción para el problema prueba *Modelo climático*.

Tipo de red neuronal	Tiempo	Porcentaje de aciertos	MSE
Multicapa	0.0132seg.	92.59	0.0741
Radial	0.0174seg.	91.48	0.0852
Probabilística	0.0192seg.	85.18	0.1418

Para la base de datos de *Plantas*, el cuadro 9 muestra los resultados del entrenamiento para cada tipo de red neuronal: cantidad de neuronas utilizadas, tiempo de entrenamiento, porcentaje de aciertos y error cuadrático medio (MSE). En el cuadro 10, se muestran los resultados de la predicción: porcentaje de aciertos, tiempo de ejecución, así como el MSE.

Cuadro 9. Resultados del entrenamiento para el problema prueba *Plantas*.

Tipo de red neuronal	Neuronas	Tiempo	Porcentaje de aciertos	MSE
Multicapa	292	45 minutos	60.375	0.0066
Radial	400	28.79seg.	99.87	0.000025
Probabilística	65	0.2058seg.	84.1250	0.0032

Cuadro 10. Resultados de la predicción para el problema prueba *Plantas*.

Tipo de red neuronal	Tiempo	Porcentaje de aciertos	MSE
Multicapa	0.0192seg.	46.875	0.0086
Radial	0.094seg.	73.0	0.0040
Probabilística	0.1840seg.	67.53	0.00658

4. Conclusiones

Las redes neuronales son técnicas excelentes que deben usarse en los problemas de clasificación debido a que sus porcentaje de clasificación son altos en la fase de entrenamiento para las bases de datos *Fertilidad, Parkinson y Modelo climático*.

La red neuronal multicapa es un modelo que tiene muchos parámetros que deben ajustarse a prueba y error, como por ejemplo, la cantidad de capas ocultas, el número de neuronas de cada capa oculta. Es difícil lograr que disminuya el error cuadrático medio, siendo por consiguiente muy tardado su entrenamiento.

La red neuronal en base radial es rápida en su entrenamiento. En su fase de entrenamiento clasifica adecuadamente para las cuatro bases de datos con al menos el 99.87% y predice con al menos el 73% de aciertos en datos que no se usan en su entrenamiento. Esto se debe a la naturaleza no lineal de su capa oculta.

La red neuronal probabilística es la más rápida, aunque resulta inferior en sus resultados cuando se compara a la red neuronal en base radial.

Se recomienda el uso de las redes neuronales en base radial, pues su funcionamiento es bastante bueno sin importar la cantidad de instancia, atributos o clases.

Para la base de datos *Plantas* la fase de prueba no proporciona porcentajes altos de clasificación correcta, la razón es que existen demasiadas clases (100) y la cantidad de instancias es bastante limitada.

Referencias

- Chen, S.L., Tsai M.T. (2014), "The Operational Monitor of Air-conditioners by using Probabilistic Neural Network", *Applied Mechanics and Materials*, 543(547), p. 1149-1152.
- Erdogdu S., Isenkul, M., Sakar, C., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., Kursun, O. (2013) "Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings", *IEEE Journal of Biomedical and Health Informatics*, vol. 17(4), pp. 828-834.
- Erdogdu, S. (2013), Parkinson Speech Dataset
<http://archive.ics.uci.edu/ml/datasets/Parkinson>
- Fecha consulta: 15 de Octubre de 2014, fecha actualización sitio web: 27Junio de 2014.
- Gila, D., Girelab J., De Juan, J., Gomez M., Johnssonc M.(2012), "Predicting seminal quality with artificial intelligence methods", *Expert systems with applications*, 39 (16).
- Gil, D y Girela L. (2013), Fertility Data Set,
<http://archive.ics.uci.edu/ml/datasets/Fertility>
- Fecha consulta: 2 de Julio de 2014, fecha actualización sitio web: 27Junio de 2014.
- Goshvarpour1, A., Ebrahimnezhad H., Goshvarpour A. (2013), "Classification of Epileptic EEG Signals using Time-Delay Neural Networks and Probabilistic Neural Networks", *I.J. Information Engineering and Electronic Business*, 1(1), p. 59-67.
- Khoshnoud, S., Ebrahimnezhad, H. (2013), "Classification of Arrhythmias Using Linear Predictive Coefficients and Probabilistic Neural Network", *Applied Medical Informatics* 33(3), p: 55 – 62.

Lucas, D. , Klein, D., Tannahill, J., Ivanova, D., Brandon, S., Domyancic, D., and Zhang, Y. (2013), “Failure analysis of parameter-induced simulation crashes in climate models”, *Geosci. Model Dev. Discuss*, 6 (6), p. 585-623.

Lucas, D. (2013), Climate Model Data Set
<http://archive.ics.uci.edu/ml/datasets/Climate+Model+Simulation+Crashes>

Fecha consulta: 2 de Juliode 2014, fecha actualización sitio web: 27Juniode 2014

Mallah, C., Orwel, J. (2013) , “Probabilistic Classification from aK-Nearest-Neighbour Classier”, *Computational Research*, 1(1), p. 12564-12573.

Mallah, C., Cope, J., Orwell, J. (2013),“Plantleaf classification using probabilistic integration ofshape, texture and margin features”,*Signal Processing, Pattern Recognition and Applications*, 1(1): 1-9

Mallah, C. (2012), 100 Leaves Plant Species Data Set
<http://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set>último acceso 2 de Julio de 2014, actualización27 de Junio de 2014.

Rojas, R. (1996), *Neural Networks a systemic introduction*, Springer-Verlag.

Tou, J.T., Gonzalez, R.C. (1974), *Pattern Recognition*, Addison-Wesley.

Capítulo 10

SECUESTRO DE CARBONO EN SUELOS DE LA RTP 105 CUETZALAN, MÉXICO, A TRAVÉS DE TELEDETECCIÓN Y MODELOS MIXTOS.

G. Linares Fleites¹; E.Torres Trejo²; M. de L. Sandoval Solís³

¹Posgrado Cs. Ambientales y Dpto. en Cs. Agrícolas

²Doctorado en Ciencias Ambientales

³Facultad de Ciencias de la Computación

Benemérita Universidad Autónoma de Puebla

Puebla, México

malosa1753@gmail.com

ABSTRACT

Carbon sequestration in soils is of great importance to the mitigation of climate change and its assessment, from the estimate of the concentration of carbon in soil, allows decision making in conservation and ecosystem management. This study is based on the digital processing of normalized vegetation index (NDVI) to estimate the degree of involvement of deforested areas through selected images coming from the satellite Landsat-5 TM and Landsat sensor 7 ETM+. The objective is to estimate, by linear mixed model, the relationship between NDVI and soil properties of the RTP region 105: Cuetzalan, defined as priority for conservation.

Keywords: climate change, vegetation indices, linear mixed regression, R Software

RESUMEN

El secuestro de carbono en suelos es de gran importancia para la mitigación del cambio climático; su evaluación, a partir de la estimación de la concentración del carbono en suelo, permite la toma de decisiones en la conservación y manejo de los ecosistemas. Este estudio está basado en el procesamiento digital del índice de vegetación normalizado (NDVI) que permite estimar los grados de afectación de áreas deforestadas a través de imágenes seleccionadas que proceden del satélite Landsat- 5 sensor TM y Landsat 7 ETM+. El objetivo es estimar, a través de modelo lineal mixto, las relaciones existentes entre NDVI y algunas propiedades del suelo de la región RTP 105: Cuetzalan, definida como prioritaria para la conservación.

1. INTRODUCCIÓN

Como resultado del aumento de concentraciones de gases de efecto de invernadero, existen evidencias científicas que sugieren que el clima global se verá alterado en este siglo. El incremento actual de la concentración de dióxido de carbono (CO₂) atmosférico, asociado al cambio climático, puede ser parcialmente mitigado por el secuestro de carbono en suelos.

El secuestro de carbono es esencialmente el proceso de transformación del carbono del aire en carbono almacenado en el suelo. El dióxido de carbono es absorbido por las plantas a través del proceso de fotosíntesis, e incorporado en la planta como tejido vegetal. Cuando las plantas mueren, el carbono de las hojas, tallos y raíces se descompone en el suelo y se convierte en materia orgánica. Este proceso reduce los niveles de CO₂ en la atmósfera, disminuyendo las probabilidades del calentamiento global.

Puede afirmarse que el mayor responsable del cambio climático global es el CO₂, que tiene entre sus fuentes emisoras la deforestación y la destrucción del suelo [1]. El carbono orgánico del suelo es un gran y activo reservorio y los ecosistemas forestales pueden absorber cantidades significativas de CO₂, por lo que hay un gran interés por incrementar el contenido de carbono en estos ecosistemas.

A pesar de la importancia del secuestro de carbono para la mitigación del cambio climático, su evaluación se encuentra muy limitada en muchas zonas del mundo, en particular, en los suelos del estado de Puebla, México.

Los métodos para estudiar los almacenes de carbono en el suelo pueden clasificarse, según el objetivo del estudio, en local, regional o global [2]. Nuestro objetivo es local y perseguimos encontrar ecuaciones que modelen el comportamiento de los contenidos de carbono orgánico en el suelo con respecto a la vegetación existente en la zona. En la actualidad se están ideando nuevas metodologías de investigación y se están utilizando nuevas herramientas de búsqueda de información. Los instrumentos de teledetección son útiles en estudios como el que presentamos, ya que permiten obtener información sobre la vegetación de la zona de interés. Esa información se puede relacionar posteriormente con las tradicionales mediciones *in situ*.

Una estrategia de mitigación llamada Reducción de Emisiones por Deforestación y Degradación de Bosques (REDD) se está desarrollando para hacer frente a las emisiones debidas a la pérdida de bosques en los países en desarrollo. REDD será el instrumento básico de cualquier acuerdo climático según el documento final de la 15^a Conferencia de las Naciones Unidas de las Partes [3]. Sin embargo, la implementación de REDD presenta retos científicos, además de los políticos. Los métodos estadísticos y, en particular los modelos mixtos, pueden jugar un papel fundamental en la evaluación de los contenidos de carbono orgánico en suelo.

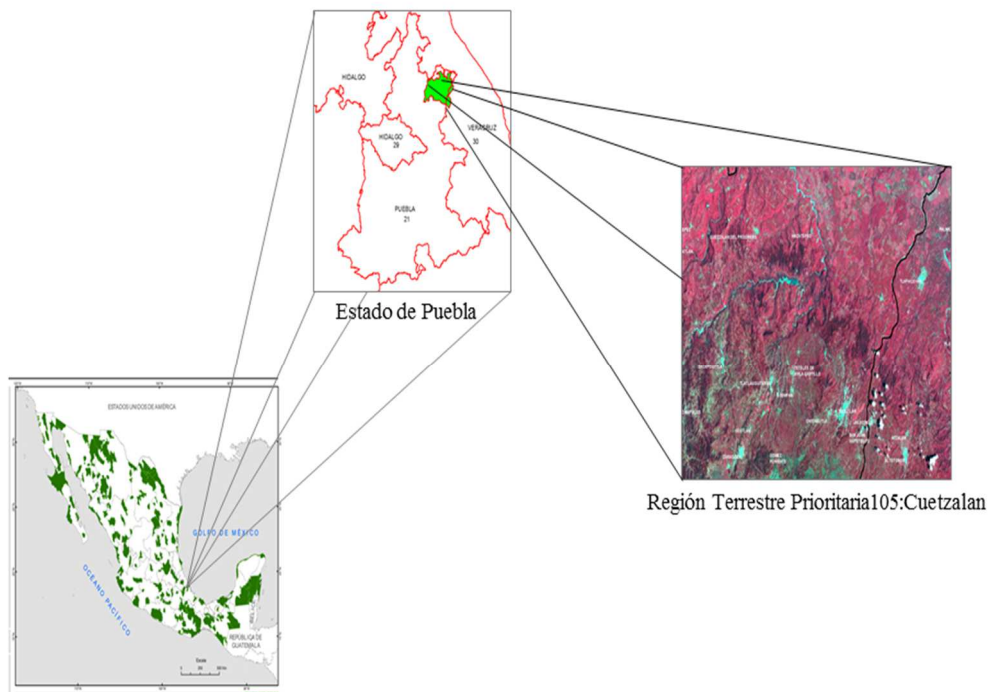
En el capítulo se hace un breve recuento de los aspectos de teledetección e índices de vegetación que pueden obtenerse a través de imágenes satelitales. Posteriormente, se relacionan los conceptos y procedimientos de los modelos lineales mixtos y se aplican para obtener las mejores ecuaciones que permitan estimar el porcentaje de carbono en suelo en la Región Terrestre Prioritaria 105, Cuetzalan, México. Finalmente, se brindan las conclusiones y las referencias.

2. TELEDETECCIÓN

El término teledetección es una traducción del inglés “*Remote Sensing*”, y se refiere, no sólo, a la captación de datos desde el aire o desde el espacio sino también a su posterior tratamiento [4]. En teledetección la adquisición de información puede realizarse a nivel orbital (satélites), suborbital (aviones) o terrestre (radiómetros portátiles). Las imágenes satelitales y las fotografías aéreas son los productos más difundidos y, en ocasiones, son

de fácil acceso y gratis. A continuación se caracteriza el área de estudio y el material satelital utilizado.

El estudio se sitúa en la Región Terrestre Prioritaria (RTP) 105. En [5] se define a una RTP como al “Área continental cuyas características físicas y biológicas la hacen particularmente importante desde el punto de vista de la biodiversidad”; son regiones que por sí mismas tienen calidad de prioritarias para preservar la biodiversidad bajo esquemas particulares de conservación, o de desarrollo sustentable. La RTP 105 se ubica en las coordenadas extremas: 19°46'23" y 20°11'55" de latitud norte y 97°09'17" a 97°38'36" de longitud oeste, con una superficie de 1,284 km² (Figura 1). Esta región se definió como prioritaria para la conservación debido a la existencia de bosques mesófilo de montaña, selva alta perennifolia en las partes bajas y bosques de encino, y bosque de encino en asociación con pino. El aprovechamiento de los recursos naturales que se realiza en la RTP se relaciona con actividades forestales, agrícolas y/o pecuarias, generalmente asociadas a la deforestación, degradación o transformación del



Regiones Terrestres Prioritarias de México.

ecosistema.

Figura 1.- Localización de la zona de la RTP 105: Cuetzalan.

El material satelital comprende: imágenes Landsat con los sensores MSS (MultiSpectral Scanner), TM (ThematicMapper), ETM+ (EnhancedThematicMapper Plus) que corresponden al satélite Landsat 5 y Landsat 7 respectivamente; la imagen TM corresponde al 3 de diciembre de 1994 y la ETM+ al 2 de noviembre del 2005, y fueron obtenidas en el sitio oficial del Servicio Geológico de los Estados Unidos (USGS, por sus siglas en inglés), <http://landsat.usgs.gov>.

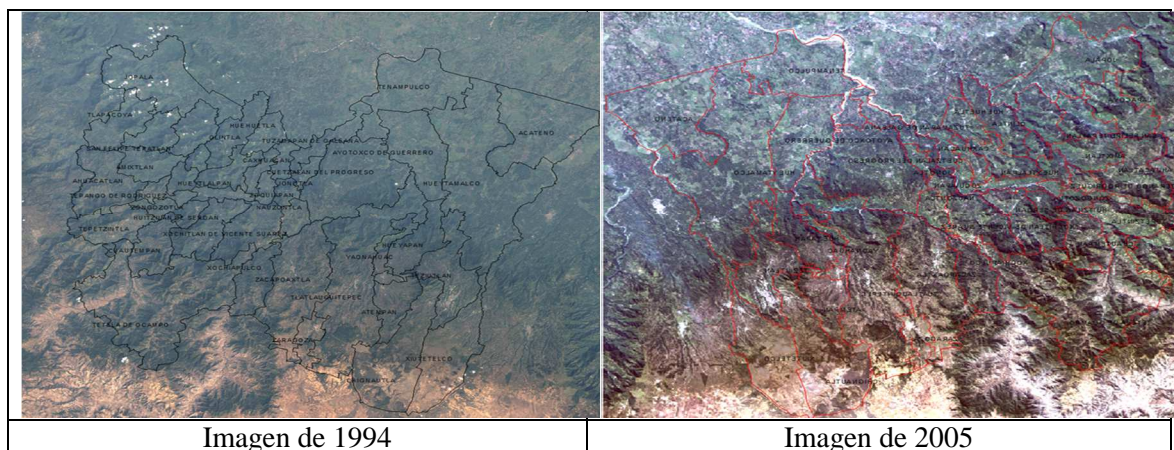


Figura 2. Imágenes de la zona de la RTP 105

A partir de estas imágenes (Figura 2) se llevó a cabo el procesamiento digital del Índice de Vegetación Normalizado (NDVI), que permite estimar el grado de afectación de áreas deforestadas, expresándolo en términos del número de píxeles.

3. ÍNDICES DE VEGETACIÓN

Los índices de vegetación son medidas cuantitativas que se basan en valores digitales útiles para estimar la biomasa o vigor vegetal. Usualmente el índice de vegetación es producto de varios valores espectrales que son sumados, divididos, o multiplicados en una forma diseñada para producir un simple valor que indique la cantidad o vigor de la vegetación dentro de un píxel. Valores altos de índices de vegetación, significan píxeles representativos para vegetación saludable; se aprovecha el comportamiento radiométrico de la vegetación sana en las diferentes bandas espectrales, especialmente entre el espectro visible y el infrarrojo cercano. En el espectro visible, en particular en la banda roja, los pigmentos de las hojas de las plantas absorben la mayor parte de la energía que reciben de sol, estas pigmentaciones apenas afectan al infrarrojo cercano [6].

Los contrastes de la imagen son utilizados para el cálculo de los índices de vegetación. La metodología se basa en el uso del Sistema de Información Geográfica (SIG) ArcMap 9.3, para el procesamiento digital e interpretación de las imágenes que proceden del satélite Landsat-5 sensor TM, que proporcionan imágenes de 8 bits en 7 y 8 bandas espectrales, tres de las cuales son visibles y se reconocen como Banda 1, Banda 2 y Banda 3 y cuatro infrarrojo, que son la Banda 4, Banda 5, Banda 6 y Banda 7. Las bandas 3 y 4, rojo visible e infrarrojo cercano, respectivamente, son adecuadas para la estimación de índices de vegetación, y facilitan establecer relaciones entre la densidad de vegetación alta y baja.

Los cálculos en este estudio corresponden al Índice de Vegetación Normalizado (NDVI). Este índice de vegetación se clasificó en dieciocho clases según su desviación estándar y, a partir de estas determinaciones, se realizó una reclasificación para obtener solo dos densidades de vegetación, la alta y la baja. La signatura espectral característica de la vegetación sana muestra un fuerte contraste entre la baja reflectividad en el rojo y la alta reflectividad en el infrarrojo. Esta diferencia es tanto mayor cuanto mayor es la densidad de la vegetación y mejor su estado fitosanitario. De esta forma para el cálculo del Índice Vegetación Normalizado (NDVI) se utilizó la siguiente ecuación:

$$X = (B4 - B3) / (B4 + B3)$$

donde:

B4: Banda 4 de la imagen de Landsat TM (reflectividad en el infrarrojo cercano)

B3: Banda 3 de la imagen de Landsat TM (reflectividad en el rojo)

Este índice varía entre -1 y 1, lo que facilita su interpretación. Puede señalarse como umbral crítico para cubiertas vegetales un valor de NDVI de 0.1 y para la vegetación densa de 0.5. Cualquier valor negativo implica ausencia de vegetación por lo que debieran reclasificarse a cero [7]

El disponer de imágenes satelitales de la región permitió estimar la densidad en la cobertura vegetal en los años 1994 y 2005 [8]. En ambos años, el NDVI se encontró en un rango de 0.1 a 0.7, que nos indica que existe una densidad de vegetación que va desde baja hasta alta en la Región Terrestre Prioritaria 105: Cuetzalan. El índice permitió identificar la presencia de vegetación verde o sana, en la superficie. (Figura 3).

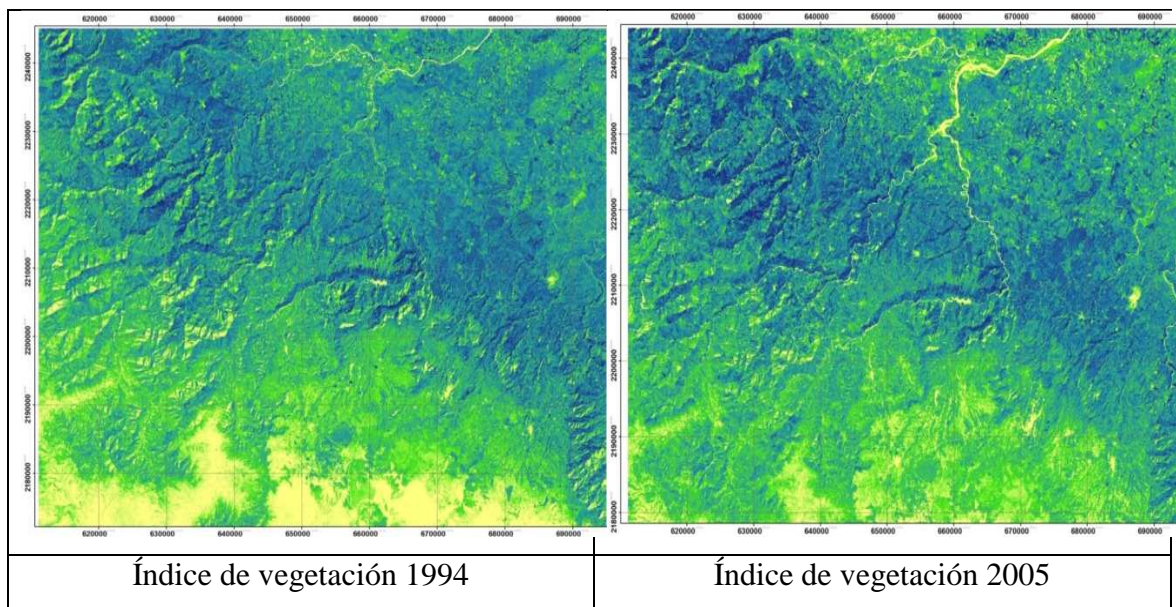


Figura 3. Imágenes de la zona de la RTP 105

4. MODELO DE REGRESIÓN LINEAL MIXTO

Los modelos mixtos son usados cuando los datos tienen algún tipo de agrupación, como es el caso que nos ocupa, donde se tiene la realización de dos niveles de densidades de vegetación (baja y alta), determinadas por el índice de vegetación NDVI. Los modelos mixtos permiten tener coeficientes fijos y aleatorios (cuyos niveles son sólo una realización de todos los posibles niveles procedentes de una población) y varios términos de error. Estos modelos pueden ser una herramienta muy útil y se cuenta actualmente con buenas referencias tales como: [9],[10], [11] y [12]

La especificación de estos modelos puede plantearse como sigue:

Sea Y_{ki} : variable dependiente continua con realizaciones y_{ki} , con i observaciones en el grupo k , donde $i = 1, \dots, n_k$ y $k = 1, \dots, K$.

- *Componente estocástico*: Y_{ki} se distribuye normal con media μ_{ki} y varianza σ^2 .
- El vector q -dimensional de efectos aleatorios, b_k , se restringe a tener media cero y por tanto se caracteriza completamente por la matriz de varianza y covarianza Ψ , simétrica y semidefinida positiva.
- *Componente sistemático*: $\mu_{ki} = X_{ki}\beta + Z_{ki}b_k$

donde

X_{ki} es el arreglo ($n_k \times p \times K$) de variables explicatorias conocidas (no aleatorias) de efectos fijos,

β es el vector p -dimensional de coeficientes de efectos fijos,

Z_{ki} es el arreglo ($n_k \times q \times K$) de variables explicatorias de efectos aleatorios conocidos, y

b_k es el vector q -dimensional de efectos aleatorios.

Para la evaluación del secuestro de carbono en la zona bajo estudio, especificaremos varios modelos mixtos en los que el efecto aleatorio se representa por los niveles bajo y alto del índice de vegetación NVDI (Indveg). Para los efectos fijos utilizaremos, según el modelo especificado, las variables densidad aparente (DA) y porcentaje de nitrógeno total (Ntot), elegidas según criterios de especialistas en edafología y estudio exploratorios previos. La variable respuesta de todos los modelos es el porcentaje de carbono en suelo (Corg). El procesamiento estadístico se realiza en lenguaje R (Bates *et al*, 2011).

En forma general, los modelos estudiados son:

Modelo 1:

$$y_{ki} = \mu + b_k + \varepsilon_{ki}$$

donde μ es la media poblacional fija y la parte $b_k + \varepsilon_{ki}$ incluye los parámetros aleatorios. Obsérvese que el error residual del modelo lineal se divide en dos partes mutuamente independientes: un efecto del grupo aleatorio para el grupo k y un residuo aleatorio para el individuo i del grupo k . Se supone que los efectos de grupos y los residuos son variables independientes y normalmente distribuidas. Este es el modelo más simple de los modelos de efectos mixtos, se le conoce como modelo de componentes de la varianza y se usa cuando no se tienen otros predictores. Los parámetros de este modelo son μ , σ_b^2 y σ^2 .

El modelo de componentes de la varianza se generaliza a situaciones donde μ se reemplaza por una función lineal de predictores. En nuestro caso tendremos el

Modelo 2:

$$y_{ki} = \beta_0 + \beta_1 X_{ki}^{(1)} + b_k + \varepsilon_{ki}$$

y el

Modelo 3:

$$y_{ki} = \beta_0 + \beta_1 x_{ki}^{(1)} + \beta_2 x_{ki}^{(2)} + b_k + \varepsilon_{ki}$$

Después de especificar un modelo con cierta parte sistemática y estructura para la parte aleatoria, el próximo paso es estimar los parámetros del modelo. Los dos métodos de estimación, más frecuentemente utilizados en la práctica, son el de mínimos cuadrados y el de máxima verosimilitud. En la literatura citada anteriormente, se recomiendan dos variantes de la máxima verosimilitud: el criterio máximo verosímil y el criterio máximo verosímil restringido, a los que nos referiremos por sus siglas en inglés como ML y REML respectivamente. El método ML puede usarse para estimar tanto los coeficientes β como los parámetros que especifican la estructura de varianza y covarianza residual, sin embargo, los estimadores de la estructura de varianza y covarianza residual son sesgados. El método REML brinda estimadores sólo para los parámetros que especifican la estructura de varianza y covarianza residual y estos estimadores son insesgados al menos en ciertos casos especiales y generalmente son menos sesgados que los estimadores [12].

El método de estimación que se utiliza en el paquete estadístico *lme4* en el lenguaje R, y que hemos usado aquí, es: primero estimar los parámetros de varianza y covarianza usando REML y posteriormente utilizar estimadores mínimos cuadrados generalizados (GLS, por sus siglas en inglés) para los parámetros de la parte sistemática. Sin embargo, debe destacarse que el método ML puro es favorable en situaciones especiales donde se requieran pruebas de hipótesis para comparar modelos.

También se utilizan los criterios de información para la comparación de modelos. Los más comúnmente utilizados son el criterio de Información de Akaike (AIC) y el criterio de información bayesiano (BIC). En la literatura consultada para estos y otros criterios de selección de modelos que aparecen en las salidas del programa *lme4*, se elige como mejor modelo aquel con los valores más pequeños del criterio.

Como es bien conocido, el ajuste de un modelo requiere de la comprobación de las suposiciones. Las suposiciones básicas de un modelo lineal son que el modelo es correcto, las observaciones son independientes y las varianzas son iguales. Además, para hacer pruebas de hipótesis sobre un modelo donde se ha utilizado estimadores mínimos cuadrados se necesita normalidad en los residuos y, en la estimación ML, la normalidad debe asumirse en la etapa de ajuste del modelo. El lenguaje R brinda herramientas para producir gráfico que permiten explorar las suposiciones del modelo.

4.1 MODELOS LINEALES MIXTOS EN 1994

La tabla de datos para ajustar modelos en este período contaba con 30 observaciones agrupadas según dos niveles del factor índice de vegetación (Indveg). Las tablas 1 y 2 muestran los estimadores REML y ML de los tres modelos analizados para la RTP 105 según la imagen satelital del año 1994.

En el modelo 1, según los estimadores REML, la media de las medias de los niveles del índice de vegetación es 5.6376 y los estimadores de los componentes de la varianza son 0.04009 para σ_b^2 y 11.38805 para la varianza de los residuos. Los estimadores ML muestran resultados similares con media de 5.652 y desviación estándar residual de 3.32.

Tabla1: Modelos Lineales Mixtos ajustados por el criterio REML ['lmerMod']

Modelo 1: Corg ~ 1 + (1 Indveg)	Modelo 2 : Corg ~ 1 + DA + (1 Indveg)	Modelo 3. Corg ~ 1 + DA + Ntot + (1 Indveg)
Convergencia REML :156.3	Convergencia REML : 135.7	Convergencia REML : 134.3
Efectos aleatorios	Efectos aleatorios	Efectos aleatorios
Grupo VarianzaDesEst	Grupo VarianzaDesEst	Grupo Varianza DesEst
Interc. 0.04009 0.2002	Interc. 0.7467 0.8641	Interc. 1.100 1.049
Residuo 11.38805	Residuo 6.5074 2.5510	Residuo 6.603 2.570
3.3746		
Numero de observaciones: 30	Numero de observaciones: 30	Numero de observaciones: 30
Grupo: Indveg con 2 niveles	Grupo: Indveg con 2 niveles	Grupo: Indveg con 2 niveles
Efectos fijos	Efectos fijos	Efectos fijos
EstimErrEst t	EstimErrEst t	EstimErrEst t
Interc 5.6376 0.6365	Interc 16.330 2.466	Interc 15.9727 2.5720
8.86	6.62	6.21
	DA -14.451 3.094 -	DA -14.3590 3.1212 -4.6
	4.67	Ntot 0.4175 0.6615 0.63
	Correlación de efectos fijos:	Correlación de efectos fijos:
	(Interc)	(Interc) DA
	DA -0.946	DA -0.923
		Ntot -0.202 0.053

En el modelo 2, la media por niveles del índice de vegetación es (16.330 -14.451 “DA”) según los estimadores REML, y (16.51 -14.37 “DA”) según los estimadores ML. Se destaca la fuerte influencia de la densidad aparente sobre el secuestro de carbono. Obsérvese que la constante del modelo 2 varía alrededor de 16.330 con una desviación estándar de 2.5510 lo que implica que el nivel de la recta varía según la densidad baja o alta del índice de vegetación. Los resultados mostrados con la estimación REML se reiteran con los estimadores ML.

En el modelo 3, los estimadores REML y ML también muestran resultados similares. La media por niveles del índice de vegetación es (15.9727 -14.3590 “DA” + 0.4175“Ntot”), según REML, y (16.3356 -14.3045 “DA” +0.2418 “Ntot”) según ML. Como en el modelo 2, se destaca la fuerte influencia de la densidad aparente sobre el secuestro d carbono y se aprecia que el efecto del nitrógeno total sobre los contenidos de carbono orgánico en el suelo no es significativo.

Teniendo en cuenta los criterios de selección de modelos, el modelo 2 puede considerarse el mejor de los tres modelos dado que los valores del AIC y del BIC son más pequeños. Sin embargo, para corroborar esta decisión, se llevó a cabo la prueba de razón de verosimilitud; debido a que el p-valor es grande (Tabla 3) no debiéramos rechazar al modelo 2 a favor del modelo 3; preferimos le modelo 2 más parsimonioso y lo consideraremos para el análisis final.

Tabla2: Modelos Lineales Mixtos ajustados por Máxima Verosimilitud (ML)
['lmerMod']

Modelo 1: Corg ~ 1 + (1 Indveg)	Modelo 2 : Corg ~ 1 + DA + (1 Indveg)	Modelo 3. Corg ~ 1 + DA + Ntot + (1 Indveg)
AIC BIC logLik 163.135 167.338 -78.5676	AIC BIC logLik 148.5451 154.1499 - 70.2726	AIC BIC logLik 150.399 157.4054 -70.1997
Efectos aleatorios Grupo DesvEst Intercepto 0.00 Residuo 3.32	Efectos aleatorios Grupo DesvEst Intercepto 0.000 Residuo 2.518	Efectos aleatorios Grupo DesvEst Intercepto 0.000 Residuo 2.512
Numero de observaciones: 30 Grupo: Indveg con 2 niveles	Numero de observaciones: 30 Grupo: Indveg con 2 niveles	Numero de observaciones: 30 Grupo: Indveg con 2 niveles
Efectos fijos Intercepto Estimador 5.652	Efectos fijos Intercepto Estimador 16.51 DA -14.37	Efectos fijos Intercepto Estimador 16.3356 DA -14.3045 Ntot 0.2418

Tabla 3: Comparación de modelos

Modelos: Modelo 2: Corg ~ 1 + DA + (1 Indveg) Modelo 3: Corg ~ 1 + DA + Ntot + (1 Indveg) Df AIC BIC Chisq Chi DfPr(>Chisq) Modelo2 : 4 148.54 154.15 Modelo3:5 150.40 157.41 0.1457 1 0.7027
--

Después de varias exploraciones para detectar datos aberrantes, decidimos tomar como ajuste del modelo el que aparece en la tabla 4 con 28 observaciones. Los gráficos que se muestran en la figura 4 permiten comprobar el cumplimiento de las suposiciones y considerar que este modelo es adecuado para predecir el contenido de carbono orgánico en suelos según la imagen satelital del año 1994 y los estudios de propiedades del suelo realizado en ese período de tiempo.

Tabla 4: Modelo mixto para estimar CO₂ en suelo. (Año1994)

Modelo 2 :				
$\text{Corg} \sim 1 + \text{DA} + (1 \text{Indveg})$				
Criterio de convergencia de REML: 103.8				
Residuos escalados:				
Min	1Q	Mediana	3Q	Max
-1.9120	-0.6466	-0.1830	0.5156	2.0618
Efectos aleatorios:				
Grupo	Varianza		Desv.Est.	
Indveg (Intercepto)	0.2274	0.4769		
Residuo	2.7751	1.6659		
Número de observaciones: 28, Grupo: Indveg, 2				
Efectos fijos:				
	Estimador	ErrorEst.	Valor de t	
(Intercepto)	14.367	1.652	8.695	
DA	-12.345	2.089	-5.909	
Correlación of Efectos Fijos:				
	(Intercepto)			
	DA	-0.957		

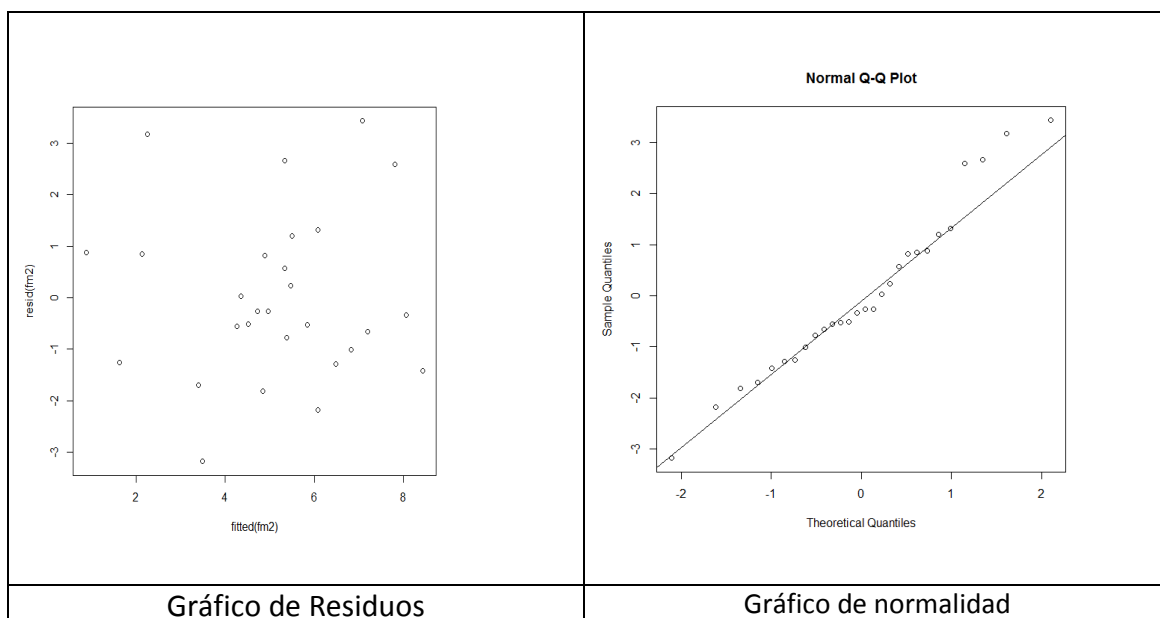


Figura 4: Gráficos para checar suposiciones (Año 1994)

4.2 MODELOS LINEALES MIXTOS EN 2005

La tabla de datos para ajustar modelos en este período contaba con 51 observaciones agrupadas según dos niveles del factor índice de vegetación (Indveg). Las tablas 5 y 6 muestran los estimadores REML y ML de los tres modelos analizados para la RTP 105 según la imagen satelital del año 2005.

En el modelo 1, según los estimadores REML, la media de las medias de los niveles del índice de vegetación es 4.1639 y los estimadores de los componentes de la varianza son cero para σ_b^2 y 10.95 para la varianza de los residuos. Los estimadores ML muestran una media de 4.164 y desviación estándar residual de 3.27, esto es, prácticamente resultados iguales.

En el modelo 2, la media por niveles del índice de vegetación es (13.404 -12.272 “DA”) según los estimadores REML, y (13.40 -12.27 “DA”) según los estimadores ML. Se destaca la fuerte influencia de la densidad aparente sobre el secuestro de carbono. Obsérvese que la constante del modelo 2 varía alrededor de 13.404 con una desviación estándar casi nula lo que implica que la recta no varía según la densidad baja o alta del índice de vegetación. Los resultados mostrados con la estimación REML se reiteran con los estimadores ML.

Tabla 5: Modelos Lineales Mixtos ajustados por el criterio REML ['lmerMod']

Modelo 1: Corg ~ 1 + (1 Indveg)	Modelo 2 : Corg ~ 1 + DA + (1 Indveg)	Modelo 3. Corg ~ 1 + DA + Ntot + (1 Indveg)
Convergencia REML : 265.5	Convergencia REML : 247.4	Convergencia REML : 242.4
Efectos aleatorios	Efectos aleatorios	Efectos aleatorios
Grupo VarianzaDesEst	Grupo VarianzaDesEst	Grupo Varianza DesEst
Interc. 0.00 0.000	Interc. 0.00 0.000	Interc. 4.742e-14 2.178e-07
Residuo 10.95 3.309	Residuo 8.439 2.905	Resid7.942e+00 2.818e+00
Numero de observaciones: 51	Numero de observaciones: 51	Numero de observaciones: 51
Grupo: Indveg con 2 niveles	Grupo: Indveg con 2 niveles	Grupo: Indveg con 2 niveles
Efectos fijos	Efectos fijos	Efectos fijos
EstimErrEst t	EstimErrEst t	EstimErrEst t
Interc4.1639 0.4634 8.986	Interc 13.404 2.354 5.693	Interc12.6722 2.3126 5.480
	DA -12.272 3.080 -3.984	DA -11.9713 2.9915 - 4.002
		Ntot1.3496 0.6691 2.017
	Correlación de efectos fijos: (Intercepto)	Correlación de efectos fijos:
	DA -0.985	(Intercepto) DA
		DA -0.979
		Ntot -0.157 0.050

En el modelo 3, los estimadores REML y ML también muestran resultados similares. La media por niveles del índice de vegetación es (12.6722 -11.9713 “DA” + 1.3496“Ntot”), según REML, y lo mismo según ML. Como en el modelo 2, se destaca la fuerte influencia de la densidad aparente sobre el secuestro d carbono. Se aprecia que el efecto del nitrógeno total sobre los contenidos de carbono orgánico en el suelo es significativo.

Tabla 6: Modelos Lineales Mixtos ajustados por Máxima Verosimilitud (ML) [lmerMod']

Modelo 1: Corg ~ 1 + (1 Indveg)	Modelo 2 : Corg ~ 1 + DA + (1 Indveg)	Modelo 3. Corg ~ 1 + DA + Ntot + (1 Indveg)
AIC BIC logLik271.7833 277.5787 - 132.8916	AIC BIC logLik259.4699 267.1972 - 125.7350	AIC BIC logLik 257.3208 266.9799 - 123.6604
Efectos aleatorios Grupo DesvEst Intercepto 0.000 Residuo 3.277	Efectos aleatorios Grupo DesvEst Intercepto 0.000 Residuo 2.848	Efectos aleatorios Grupo DesvEst Intercepto 0.000 Residuo 2.734
Numero de observaciones: 51 Grupo: Indveg con 2 niveles	Numero de observaciones: 51 Grupo: Indveg con 2 niveles	Numero de observaciones: 51 Grupo: Indveg con 2 niveles
Efectos fijos Estimador Intercepto 4.164	Efectos fijos Estimador Intercepto 13.40 DA -12.27	Efectos fijos Estimador Intercepto12.67 DA -11.97 Ntot1.35

Teniendo en cuenta los criterios de selección de modelos, el modelo 3 puede considerarse el mejor de los tres modelos dado que los valores del AIC y del BIC son más pequeños. Para corroborar esta decisión, se llevó a cabo la prueba de razón de verosimilitud (Tabla 7); debido a que el p-valor es pequeño (menor a un nivel de significación $\alpha = 0.05$) se debe rechazar al modelo 2 a favor del modelo 3. El modelo 3 más será considerado para el ajuste final.

Tabla 7: Comparación de modelos

Modelos: Modelo 2: Corg ~ 1 + DA + (1 Indveg) Modelo 3: Corg ~ 1 + DA + Ntot + (1 Indveg) DfAIC BIClogLik Chi DfPr(>Chisq) Modelo 2: 4 259.47 267.20 -125.73 Modelo 3: 5 257.32 266.98 -123.66 1 0.04166 *
--

Después de varias exploraciones para detectar datos aberrantes, decidimos tomar como ajuste del modelo el que aparece en la tabla 8 con 40 observaciones. Los gráficos que se muestran en la figura 5 permiten comprobar el cumplimiento de las suposiciones y considerar que este modelo es adecuado para predecir el contenido de carbono orgánico en suelos según la imagen satelital del año 2005 y los análisis de propiedades del suelo realizado en ese período de tiempo.

Tabla 8. Modelo mixto para estimar CO₂ en suelo. (Año2005)

Modelo 3 :			
$Corg \sim 1 + DA + Ntot + (1 Indveg)$			
Criterio de convergencia de REML: 155.9			
Efectos aleatorios:			
Grupo		Variance	Std.Dev.
Indveg (Intercepto)	0.2423	0.2423	0.4922
Residuo	3.3017	3.3017	1.8170
Número de observaciones: 40			
Grupo: Indveg, 2			
Efectos fijos:			
	Estimador	ErrorEst.	Valor de t
(Intercepto)	7.2509	1.9512	3.716
DA	-5.6929	2.3768	-2.395
Ntot 0.7968	0.4443	1.793	
Correlación of Efectos Fijos:			
(Intr) DA			
DA	-0.968		
Ntot	-0.096	0.021	

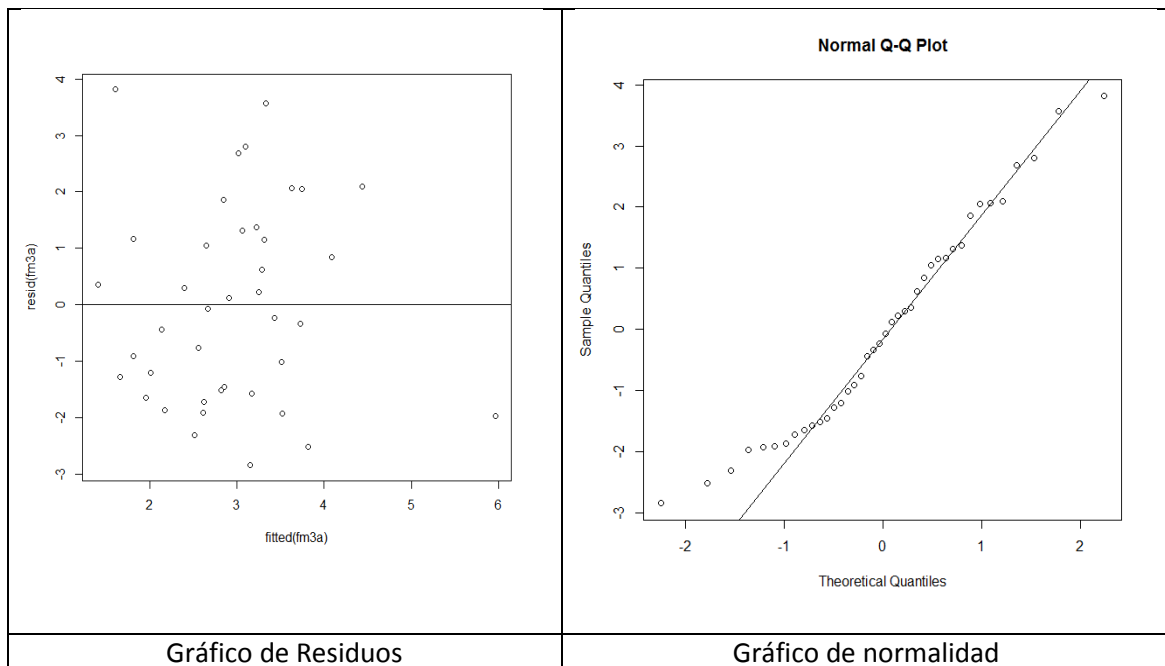


Figura 5: Gráficos para checar suposiciones (Año 2005).

5. CONCLUSIONES

El estudio de índices de vegetación en la Región Terrestre Prioritaria 105: Cuetzalan, Puebla, México, mediante el uso de sensores remotos, permitió contar con valores del índice de vegetación NDVI, y con ellos, visualizar el contraste entre dos clases de densidad de vegetación, la baja y la alta. Bajo la consideración que estos dos niveles son una muestra de la población de niveles del índice NDVI, se procedió a ajustar diferentes modelos de regresión lineal mixta.

Se identificaron dos modelos de regresión mixtos, uno por cada imagen satelital, que pueden ser útiles en la evaluación del carbono orgánico en suelo en la región considerada. Para la imagen del año 1994, el modelo estuvo en función de un único predictor (la densidad aparente de esos suelos), mientras que para la imagen del año 2005, el modelo con dos predictores (densidad aparente y porcentaje de nitrógeno) mostró las mejores características.

Profundizar en la modelación estadística para obtener buenas estimaciones y predicciones del carbono orgánico en suelos, es de suma importancia para aumentar el conocimiento sobre esta forma de mitigación del cambio climático: el secuestro de carbono.

REFERENCIAS

- [1] MASERA, O. and SHEINBAUM, C. (2004). **Mitigación de emisiones de carbono y prioridades de desarrollo nacional.** En: *Cambio climático: una visión desde México.* Compiladores: Martínez, J. y Fernández, A. Instituto Nacional de Ecología. Secretaría del Medio Ambiente y Recursos Naturales. México, pp. 255-368.
- [2] RODEGHIRO, M, HEINEMEYER, A., BELLAMY, P. (2010). **Determination of soil carbon stocks and changes.** En: *Soil Carbon Dynamics. An Integrated Methodology.* Kutsch, W. L. Bahn, M. and Heinemeyer, A. Cambridge University Press, UK, 49-75.
- [3] ARCIDIACONO-BÁRSONYA, C., CIAISA, P., VIOVYA, N. VUICHARD, N. (2011). *REDD Mitigation.* Elsevier BV. **Procedia Environmental Sciences** 6. 50 – 59.
- [4] ZHENG, D.; RADEMACHER, J.; CHEN, J.; CROW, T.; BRESEE, M.; LE MOINE, J. y RYU, S. (2004). *Estimating aboveground biomass using Landsat 7 ETM+ data across a managed landscape in northern Wisconsin, USA.* **Remote Sensing of Environment.** v.93, p.402– 411.
- [5] ARRIAGA, L.; J.M. ESPINOZA; C. AGUILAR; E. MARTÍNEZ; L. GÓMEZ y E. LOA (Eds). (2000). **Regiones Terrestres Prioritarias de México. Comisión Nacional para el conocimiento y uso de la biodiversidad,** México.
- [6] CHUVIECO, E. (2007). *Earth Observation and Global Change.* Nueva York: Springer. Geosciences.
- [7] ROUSE, J.W., HAAS, R.H., SCHELL, J.A., DEERING, D.W. and HARLAN, J.C., (1974). Monitoring the vernal advancement of retrogradation of natural vegetation. NASA/GSFC, Type III, Final Report, Greenbelt, MD, USA, 371 p.
- [8] TORRES, E.; LINARES, G.; TENORIO, M.G.; PEÑA, R.; CASTELÁN, R.; RODRÍGUEZ, A. (2014). *Obtención de índices de vegetación para estimar densidades de vegetación y cambios*

de Uso de Suelo en la Región Terrestre Prioritaria 105: Cuetzalan, Puebla, México.
ACEPTADO en **Revista Iberoamericana de Ciencias**. ISSN 2334-2501.

[9] PINHEIRO, J.C. and BATES, D.M. (2000). *Mixed-effects Models in S and S-PLUS*. New York: Springer-Verlag. 528 p

[10] DEMIDENKO, E.(2013). *Mixedmodels.Theory and applications with R*.2nd edition.Wiley, New York, USA.

[11] GALECKI, A., and BURZYKOWSKI, T. (2013). *Linear mixed-effects models using R.A step by step approach*.1st edition.Springer, New York, 542 p.

[12] MEHTÄTALO, L. (2013). *Linear mixed-effects models with examples in R*. University of Eastern Finland. 158 p.

[13] BATES, D., MAECHLER, M and BOLKER, B.(2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-42.

Capítulo 11

MULTICLASIFICADOR PARA DIAGNOSTICAR HIPERTENSIÓN ARTERIAL INFANTIL COMBINANDO ALGORITMOS GENÉTICOS CON MEDIDAS DE DIVERSIDAD

L. Cabrera Hernández, L. Denoda Pérez, A. Morales Hernández, G.M. Casas Cardoso, E. González Rodríguez
Universidad Central “Marta Abreu” de Las Villas
{leidysc, ldenoda, alejandrom, gcasas, eglez}@uclv.edu.cu

ABSTRACT

Nowadays, deaths by cardiovascular diseases are increasing. It is known that High Blood Pressure produces an increase the probability to suffer a cardiovascular disease. The early diagnostic of hypertension is consequently a very important problem. In this paper a muticlassifier to diagnostic high blood pressure in kids of 10 to 12 years are presented. Genetic Algorithm Metaheuristic is used to combine the information of 18 base classifiers with 13 diversity measures and to obtain a good enough multiclassifier. The accuracy of the new classifier is superior to the accuracies of all base classifiers.

Key Words: multiclassifier, diversity measures, genetic algorithm, high blood pressure in kids

RESUMEN

En la actualidad, las muertes por enfermedades cardiovasculares van en ascenso. Se conoce que la elevada presión arterial incrementa la probabilidad de padecer una enfermedad cardiovascular. El diagnóstico temprano de la hipertensión arterial es un problema muy importante. En este artículo se presenta un multclasificador para diagnosticar hipertensión en niños de 10 a 12 años de edad. Se utiliza la metaheurística basada en algoritmos genéticos para combinar la información de 18 clasificadores bases con 13 medidas de diversidad y así obtener un buen multclasificador. La exactitud de este nuevo clasificador es superior a la de los clasificadores individuales.

Palabras claves: multclasificador, medidas de diversidad, algoritmos genéticos, hipertensión arterial en niños

1. INTRODUCCIÓN

El término de hipertensión arterial es cada vez más común en nuestra sociedad y su identificación como factor de riesgo cardiovascular. A pesar de ello, no todo el mundo traslada esta preocupación a los niños. La prevención de las enfermedades cardiovasculares no queda limitada a la edad adulta, sino que debe iniciarse en la edad pediátrica.

Las técnicas de clasificación están siendo muy utilizadas en la solución de diferentes problemas de la sociedad. Existen varios modelos de clasificación reportados en la literatura como las redes neuronales, árboles de clasificación y análisis discriminante. En investigaciones recientes muchos autores introducen el término multclasificador como un “clasificador” que combina las salidas de un conjunto de clasificadores individuales, utilizando algún criterio (ej.; promedio, voto mayoritario, mínimo, etc.). Cuando se combinan clasificadores es importante garantizar la diversidad entre ellos ya que no tendría sentido combinar clasificadores cuya clasificación sea la misma. Existen varios modelos para construir un multclasificador y todos garantizan esta diversidad de diferentes formas. En el caso de aquellos que usan distintos clasificadores bases, existen algunas medidas estadísticas que pueden ser usadas para estimar cuán diversos son. Muchos estudios teóricos y empíricos han sido publicados demostrando las ventajas del paradigma de combinación de clasificadores sobre el de clasificadores individuales.

Además de los modelos tradicionales de construcción de multclasificadores recientemente se ha comenzado con la utilización de varias meta-heurísticas combinadas con este tema pues debido a la gran cantidad de clasificadores individuales existentes se hace cada vez más difícil escoger cuál de ellos combinar, dentro de las meta-heurísticas existentes se destacan los Algoritmos Genéticos por su capacidad de adaptabilidad a varios problemas, específicamente al nuestro.

En este trabajo se presentan algunas medidas de diversidad y se implementa una variante de Algoritmo Genético con el objetivo de obtener una combinación de clasificadores diversos y una exactitud del sistema multclasificador superior a la mejor exactitud individual. Además se muestra una aplicación del algoritmo propuesto para predecir el riesgo de hipertensión arterial en niños.

El siguiente epígrafe describe los métodos de clasificación más comunes. Posteriormente se enuncian y describen las medidas de diversidad usadas en este trabajo, luego se presentan algunos conceptos básicos de algoritmos genéticos y su modelación para resolver el problema. Para finalizar se muestra una aplicación real con escolares de la ciudad de Santa Clara, Cuba para predecir el riesgo de hipertensión arterial en la edad pediátrica. El trabajo culmina con conclusiones y referencias bibliográficas.

2. MÉTODOS DE CLASIFICACIÓN

Los métodos matemáticos de clasificación están caracterizados fundamentalmente porque se conoce la información acerca de la clase a la que pertenece cada uno de los objetos. Cuando la variable de decisión, función o hipótesis a predecir es continua, a los algoritmos relacionados con los problemas supervisados se les conoce como métodos de regresión. Si por el contrario la variable de decisión, función o hipótesis es discreta, ellos se conocen como métodos de clasificación o simplemente clasificadores.

En un problema de clasificación se tienen un conjunto de objetos, elementos, instancias u observaciones divididos en clases o etiquetados. Dado un elemento del conjunto, un especialista le asigna una clase de acuerdo a los rasgos, características o variables que lo

describen. Esta relación entre los descriptores y la clase puede estar dada por un conjunto de reglas. La mayoría de las veces este conjunto de reglas no se conoce y la única información que se tiene es el conjunto de ejemplos etiquetados, de forma tal que las etiquetas representan las clases.

De manera general, se puede decir que los métodos de clasificación son un mecanismo de aprendizaje, donde la tarea es tomar cada instancia y asignarla a una clase en particular.

La clasificación puede dividirse en tres procesos fundamentales: pre-procesamiento de los datos, selección del modelo de clasificación y, entrenamiento y prueba del clasificador [1].

Entre los métodos de clasificación más usados están los algoritmos basados en casos, los árboles de decisión, las redes bayesianas, las redes neuronales artificiales, el análisis discriminante y la regresión logística, pero estos no son los únicos. A continuación se presenta una breve descripción de los mencionados anteriormente.

Algoritmos basados en casos

El razonamiento basado en casos se basa en el principio de usar experiencias viejas para resolver problemas nuevos. Muchos algoritmos usan este razonamiento para resolver los problemas y entre los más comunes están los de clasificación.

Aunque todos los métodos de clasificación se basan en casos, existe un conjunto que se conoce como algoritmos basados en casos, o también como métodos de aprendizaje perezoso. Estos algoritmos deben contar con una serie de ejemplos ya conocidos y cuando van a resolver un problema nuevo, lo hacen buscando la semejanza entre éste y los ejemplos almacenados. No necesitan crear reglas, ni árboles, ni ajustar parámetros. A cada ejemplo se le conoce como instancia y a la colección de ejemplos como base de casos.

Una nueva instancia se compara con el resto de la base de casos a través de una medida de similitud. La clase de la nueva instancia será la misma que la del caso que más cercano esté a la nueva instancia. A este proceso se le conoce con el nombre de método del “vecino más cercano” (nearest neighbor). Si en lugar de usar el caso más cercano se utilizan los k casos más similares, entonces se habla de los k -vecinos más cercanos¹ y la clase asignada a la nueva instancia será la más común entre las k instancias más cercanas encontradas en la base de casos [2].

Árboles de decisión

Un árbol de decisión es un modelo matemático que clasifica las instancias ordenándolas de la raíz a las hojas. Cada nodo interior del árbol especifica una prueba de algún atributo o variable y las hojas son las clases en las cuales se clasifican las instancias u observaciones. Cada rama descendiente de un nodo interior corresponde a un valor

¹ kNN por sus siglas en inglés (k Nearest Neighbors). Conocido además como IBk (IB1 cuando el número de vecinos es uno) en la plataforma inteligente para aprendizaje *Waikato Environment for Knowledge Analysis-WEKA*.

posible del atributo probado en ese nodo. Un árbol de decisión representa una disyunción de conjunciones sobre los valores de los atributos. Así, cada rama, de la raíz a un nodo hoja, corresponde a una conjunción de atributos y el árbol en sí, a una disyunción de estas conjunciones.

La familia de algoritmos ID3 [3] es el paradigma de los métodos para descubrir reglas usando árboles de decisión; a pesar de esto, tiene algunas limitaciones. Una variante para la solución de estas limitaciones es el algoritmo C4.5 [4]², que usa puntos de corte e introduce varias medidas para evitar el sobre entrenamiento, en particular los criterios de parada de la división y de poda del árbol.

Otros árboles de decisión son el CHAID (Chi Square Automatic Interaction Detector) en el que la segmentación ocurre siguiendo criterios chi-cuadrados y el CRT (Classification and Regression Tree) en el que se dividen los casos en segmentos que son lo más homogéneos posibles con respecto a la variable dependiente. Varios de estos árboles se pueden encontrar en WEKA; por ejemplo: J48, Id3, BFTree, NBTree, entre otros.

Redes bayesianas

Una red bayesiana es un modelo gráfico probabilístico que representa un conjunto de variables y sus dependencias probabilísticas. Las redes bayesianas permiten declarar supuestos de independencia condicionales que son aplicados a subconjuntos de variables. Son representadas por un gráfico acíclico dirigido, donde cada variable se representa por un nodo de la red, y de ella se especifican dos tipos de información:

- la estructura de dependencias condicionales que son los arcos de la red
- las distribuciones de probabilidad correspondientes.

Una red bayesiana puede calcular la distribución de probabilidad para cualquier subconjunto de variables de la red, dado los valores o distribuciones de las variables restantes [2]. Cuando no se conocen todos los valores de las variables en el conjunto de entrenamiento, el aprendizaje con una red bayesiana puede ser más difícil.

Este tipo de clasificador no es muy sensible a los cambios de sus parámetros, ya que se basa en información de toda la base, lo cual hace que pequeños cambios en la base no sean necesariamente significativos [5]. En WEKA hay varias de estas redes implementadas, las más sobresalientes son NaiveBayes y sus variantes.

Redes neuronales artificiales

Una red neuronal es un modelo computacional que pretende simular el funcionamiento del cerebro a partir del desarrollo de una arquitectura que toma rasgos del funcionamiento de las neuronas sin llegar a desarrollar una réplica del mismo [6]. Es una herramienta matemática para la modelación de problemas, que permite obtener las relaciones funcionales subyacentes entre los datos involucrados en problemas de clasificación, reconocimiento de patrones, regresión, etc. Este tipo de método se considera como un excelente aproximador de funciones, esencialmente no lineales,

² Conocido como ADTree en WEKA.

siendo capaces de aprender las características relevantes de un conjunto de datos, para luego reproducirlas en entornos ruidosos o incompletos [7].

En los últimos años se han producido una amplia variedad de arquitecturas de redes neuronales, encontrándose entre las más utilizadas, las redes multicapa de alimentación hacia adelante (Feed-Forward Neuronal Networks, FFN), las cuales se distinguen porque sus neuronas están conectadas a manera de grafo acíclico dirigido (todos los arcos hacia adelante). Las redes Multi Layer Perceptron (MLP) constituyen un ejemplo genérico de las redes FFN, y se encuentran formadas por un conjunto de capas de neuronas ordenadas secuencialmente. Primero una capa de entrada, luego un conjunto de capas intermedias denominadas capas ocultas y por último una capa de salida. Las MLP usando neuronas ocultas con funciones no lineales, son capaces de aproximar cualquier tipo de función continua y brindar excelentes resultados en las tareas de clasificación [8].

Regresión logística

La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia se ha puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas o categóricas. En este último caso, se requiere que sean transformadas en variables “*dummy*”; es decir, variables simuladas.

El propósito del análisis consiste en predecir la probabilidad de que a alguien le ocurra cierto “evento”. Puede, además, determinar cuáles variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión. Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso de cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos.

La regresión logística sólo resuelve problemas de clasificación binarios. Si el problema fuese más general, entonces se puede aplicar un modelo más general basado en los mismos principios, denominado regresión multinomial, precisamente este criterio es el que utiliza la función Logistic, implementada en WEKA.

Como se ha visto, se han desarrollado un gran número de clasificadores, pero determinar cuál de ellos logra encontrar una mejor frontera de decisión para separar las clases es el mayor problema. En la búsqueda de mejores métodos de clasificación aparece una tendencia a combinar varios de estos clasificadores. Los algoritmos llamados multclasificadores se basan en esta idea; utilizar varios clasificadores y combinar sus diferentes salidas [9] con el objetivo de alcanzar un mejor resultado.

3. MEDIDAS DE DIVERSIDAD COMO CRITERIO PARA SELECCIONAR LOS CLASIFICADORES DE BASE

Existen dos tipos de medidas de diversidad: las medidas pareadas (pairwise) y las medidas grupales (non pairwise).

Las medidas en forma de pares se calculan por pares de clasificadores usando sus salidas, las cuales son binarias (1,0) que indica si la instancia fue correctamente clasificada o no por el clasificador.

A continuación se indica en la Tabla 1 el resultado de dos clasificadores (C_i , C_j) para una instancia en cuanto a si la clasificaron correctamente o no.

Tabla 1: Matriz binaria para una instancia

	C_j correcto (1)	C_j incorrecto (0)
C_i correcto (1)	a	B
C_i incorrecto (0)	c	D
$a + b + c + d = 1$		

Si se suman para todas las instancias los valores de a, b, c, d se obtendrán los resultados mostrados en la Tabla 2:

Tabla 2: Matriz binaria para N instancias

	C_j correcto (1)	C_j incorrecto (0)
C_i correcto (1)	A	B
C_i incorrecto (0)	C	D
$A + B + C + D = N$		

Donde N es el número total de instancias. Un conjunto de L clasificadores produce $L(L - 1)/2$ pares de valores. Para obtener un único resultado habría que promediar estos valores.

A continuación se presentan las medidas pareadas tenidas en cuenta en este trabajo.

Coefficiente de correlación ρ

Entre las medidas de diversidad está el coeficiente de correlación [10] el cual se calcula como,

$$\rho_{c_i, c_j} = \frac{A \times D - B \times C}{\sqrt{(A+B) \times (C+D) \times (A+C) \times (B+D)}} \quad (1)$$

Mientras menor sea el valor de ρ , mayor será la diversidad. Los valores de ρ estarán en el intervalo $[-1, 1]$.

El estadístico Q

El estadístico Q (Q Statistics) es otra de las medidas para pares de clasificadores. Se calcula de la siguiente forma:

$$Q_{c_i, c_j} = \frac{A \times D - B \times C}{A \times D + B \times C} \quad (2)$$

Para cualquier par de clasificadores, los valores de ρ y Q tendrán el mismo signo y se puede probar que $|\rho| \leq |Q|$ [11].

La medida de diferencias

La medida de diferencias (The Disagreement Measure) introducida por Skalak [12], es igual a la probabilidad de que los dos clasificadores discrepen en sus predicciones. Mientras mayor sea su valor mayor será la diversidad.

$$D_{ci,cj} = \frac{B+C}{N} \quad (3)$$

La medida de doble fallo

La medida de doble fallo (The Double- Fault Measure) introducida por Giacinto y Roli [13] considera el fallo de los dos clasificadores al mismo tiempo. Está basada en el concepto de que es más importante conocer cuando errores simultáneos son cometidos que cuando ambos tienen clasificación correcta. Mientras menor sea el valor mayor será la diversidad.

$$D_{ci,cj} = \frac{D}{N} \quad (4)$$

A continuación se enuncian las medidas grupales utilizadas en este trabajo.

La medida de Entropía

La medida de Entropía (The Entropy Measure) [11] se basa en la idea intuitiva de que en un conjunto de N casos y L clasificadores la mayor diversidad se obtendrá si $L/2$ de los clasificadores clasifican una instancia correctamente y los otros $L - L/2$ la clasifican incorrectamente. Fue introducida por Cunningham y Carney en [14].

$$E = \frac{1}{N} \cdot \frac{2}{L-1} \sum_{j=1}^N \min\{(\sum_{i=1}^L y_{j,i}), (L - \sum_{i=1}^L y_{j,i})\} , y_{j,i} \in \{0,1\}, 0 \leq E \leq 1 \quad (5)$$

Varianza de Kohavi-Wolpert

La varianza de Kohavi-Wolpert (Kohavi-Wolpert Variance), fue propuesta en [15]. Esta medida es originada de la descomposición de la varianza del sesgo del error de un clasificador.

$$KW = \frac{1}{NL^2} \sum_{j=1}^N Y(z_j) (L - Y(z_j)) , 0 \leq KW \leq 1 \text{ donde } Y(z_j) = \sum_{i=1}^L y_{i,j} \quad (6)$$

Con esta medida, la diversidad disminuye a medida que el valor de KW aumenta.

Medida de desacuerdo entre expertos

La medida de desacuerdo entre expertos (Measurement interrater agreement) [16]. Se desarrolla como una medida de fiabilidad entre clasificadores. La diversidad disminuye cuando el valor de k aumenta. El k se calcula por:

$$k = 1 - \frac{\frac{1}{L} \sum_{j=1}^N Y(Z_j) (L - Y(Z_j))}{N(L-1)p(1-p)}, -1 \leq k \leq 1 \quad (7)$$

Donde el término de la derecha es la medida de concordancia de Kendall y p es la media de la exactitud de la clasificación individual, y se calcula como:

$$p = \frac{1}{N \cdot L} \cdot \sum_{j=1}^N \sum_{i=1}^L y_{j,i} \quad (8)$$

Medida de dificultad

La medida de dificultad (The Measure of "difficulty" θ) viene del estudio realizado en [17]. Se calcula a través de la varianza de una variable aleatoria discreta y denota la probabilidad de que exactamente i clasificadores hayan clasificado bien todas las instancias. La diversidad del ensamblado aumenta con el decremento del valor de la medida de dificultad.

$$\theta = Var(x) \quad (9)$$

Medida de diversidad generalizada

La medida de diversidad generalizada (Generalized Diversity) se enunció en [18].

$$GD = 1 - \frac{p(2)}{p(1)}, 0 \leq GD \leq 1, \quad \text{donde} \quad (10)$$

$$p(1) = \sum_{i=1}^L \frac{i}{L} * p[i], \quad p(2) = \sum_{i=1}^L \frac{i * (i-1)}{L * (L-1)} * p[i] \quad (11)$$

El valor de GD varía entre 0 y 1, siendo 0 la menor diversidad cuando $p(2)=p(1)$ y 1 la mayor diversidad cuando $p(2)=0$ y L la cantidad de clasificadores.

Medida de diversidad de coincidencia de fallos

Esta medida (Coincident Failure Diversity) se enuncia también en [18], como una mejora a la medida anterior.

$$CFD = \begin{cases} 0, & p[0] = 1 \\ \frac{1}{1-p[0]} * \sum_{i=1}^L \frac{L-i}{L-1} * p[i], & p[0] < 1 \end{cases} \quad (12)$$

El valor de CFD está entre 0 y 1 y mientras mayor sea, mayor será la diversidad.

Medida de diversidad de distintos fallos

Esta medida (Distinct Failure Diversity) fue igualmente enunciada en [18], como una mejora a la medida anterior.

$$DFD = \begin{cases} 0, & t[i] = 0 \\ \sum_{i=1}^L \frac{L-i}{L-1} \times t[i], & t[i] > 0 \end{cases} \quad (13)$$

El valor de DFD está entre 0 y 1 y mientras mayor sea, mayor será la diversidad.

Medida de la diversidad global

La medida de la diversidad global (Overall Diversity) fue enunciada en [19] como una versión “*pesada*” de la medida de diversidad de distintos fallos.

$$DFD = \begin{cases} 0, & t[i] = 0 \\ \sum_{i=1}^L \frac{L-i}{L-1} \times t[i] \times w[i], & t[i] > 0 \end{cases} \quad (14)$$

Cada posición de w representa el promedio de valores d para cada fila donde i clasificadores fallaron. Los valores d se calculan para cada instancia como:

$$d_i = \sum_{j=0}^K \left[\sqrt{\frac{C_{ki}}{n_i^2}} \right] \quad (15)$$

El valor de OD está entre 0 y 1 y mientras mayor sea, mayor será la diversidad.

Medida de variabilidad

Esta medida (The Measure of Variability) tiene en cuenta si las clases asignadas por los clasificadores en cada instancia son distintas o no. Mientras mayor sea su valor, mayor es la diversidad.

$$Var = \frac{\sum_{i=1}^N a}{N} \text{ donde } a = \begin{cases} 0 & \text{si } E_1(i) = E_2(i) = \dots = E_L(i) \\ 1 & \text{e. o. c} \end{cases} \quad (16)$$

Para lograr combinar varias medidas de diversidad y obtener un solo valor se tiene en cuenta el operador **Fuzzy**. Este utiliza la teoría de los conjuntos borrosos para calcular el promedio de pertenencia de cada una de las medidas a los conjuntos borrosos y retorna el máximo estandarizado de esos valores. Los términos lingüísticos manejados por este operador son *baja diversidad* y *alta diversidad*.

Las funciones de pertenencia usadas fueron triangulares y su formulación se define en la figura 1.

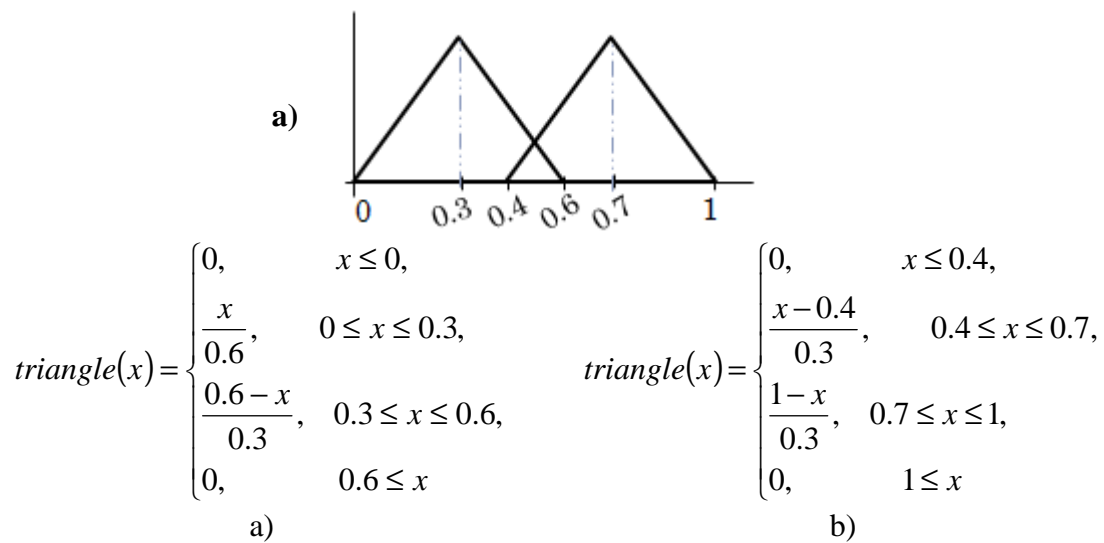


Figura 1: Funciones de pertenencia para los términos lingüísticos *baja diversidad* y *alta diversidad*. a) Representación gráfica, b) y c) Representación analítica de estos términos

Con lo anterior se puede determinar la pertenencia a cada uno de los conjuntos de las medidas de diversidad que se estén analizando.

4. ALGORITMOS GENÉTICOS COMBINADO CON MEDIDAS DE DIVERSIDAD PARA OBTENER UN BUEN MULTICLASIFICADOR

Los algoritmos genéticos (AGs) son métodos de búsqueda basados en los principios generales de la genética natural, se ejecuta para un número fijo de generaciones o hasta que algún criterio de parada es satisfecho.

La configuración del Algoritmo Genético depende del tipo de problema a resolver. En este caso, el algoritmo genético es presentado usando medidas de diversidad con el objetivo de combinar clasificadores diversos y proveer la mejor exactitud posible. El conjunto de todos los parámetros del algoritmo genético y la definición de la función objetivo son:

Configuración del cromosoma

El cromosoma representará las posibles soluciones de nuestro problema.

Gen: variable binaria que toma valor 1 si el clasificador pertenece a la combinación y 0 en otro caso.

Cromosoma: Secuencia de genes que representan el conjunto de todos los clasificadores base que serán usados en el sistema multclasificador.

La siguiente ecuación muestra los aspectos anteriores:

$$C_x = (g_1, g_2, \dots, g_L) \quad g_i = \begin{cases} 0 & , \text{clasificador } i \text{ no está presente} \\ 1 & , \text{clasificador } i \text{ está presente} \end{cases}$$

Descripción de la función objetivo

Como se desea obtener de forma simultánea la mejor diversidad entre los clasificadores que son usados en el sistema multclasificador y la mejor exactitud que se pueda obtener con él, se calcula el valor de f que es la suma entre la exactitud del sistema multclasificador y el resultado de las medidas de diversidad, según la configuración del cromosoma.

$$f(C_x) = \text{Exactitud}(C_x) + \text{Diversidad}(C_x), \quad \text{donde } C_x \text{ es el cromosoma}$$

Por tanto, la función objetivo en el proceso evolucionario será:

$$\max_{0 \leq x \leq P} f(C_x), \quad \text{donde } P \text{ es el tamaño de la población}$$

Dado que pueden existir casos en los cuales el primer parámetro de la función objetivo podría ser pequeño y el valor de $f(C_x)$ ser alto porque hay una gran diversidad entre los clasificadores. Como el principal objetivo en este trabajo es encontrar la combinación con exactitud superior en la clasificación y al mismo tiempo hallar la diversidad superior entre los clasificadores, entonces se agrega otra restricción, la cual manifiesta que el resultado final será la combinación donde la **exactitud del multclasificador** sobrepase la mejor exactitud obtenida con los clasificadores individuales y entre ellas, la combinación con más **diversidad**.

Configuración de la población

En la configuración de la población varios elementos son necesarios; por ejemplo, el número de individuos en la población y el número de ellos que serán reemplazados en cada iteración.

Se utiliza un tamaño de población, igual a $\frac{L}{2^2}$ donde L es el número de clasificadores; es decir, el tamaño de los cromosomas. Este tamaño fue sugerido tratando de evitar pequeños espacios de soluciones o un alto tiempo computacional mientras se analiza este espacio.

La población inicial será generada usando un híbrido entre la generación aleatoria y el sembrado de individuos.

Cada cromosoma es generado aleatoriamente y el valor del gen es 0 o 1 dependiendo de la presencia del clasificador en la combinación, para ello un número aleatorio r es generado, si r es mayor que 0.5 el clasificador será incluido y por eso el gen será 1; en otro caso, el clasificador no es incluido y el gen será 0. Después de que todos los cromosomas son generados como se explica antes, los mejores clasificadores individuales también serán incluidos en la combinación, poniendo el valor correspondiente de este gen igual a 1.

Los operadores de selección, cruzamiento y mutación son explicados a continuación, ellos son usados para simular la recombinación genética y el mecanismo de selección natural.

Operador de cruzamiento

En el caso del cruzamiento, es permitido seleccionar fragmentos del genotipo de cromosomas que no son muy buenos independientemente, pero cuando son mezclados, pueden ser una mejor solución respecto a la anterior. Hay varias formas de definir este operador; en nuestro caso, se usó el operador clásico de cruzamiento en un punto y el cruzamiento uniforme.

En el cruzamiento en un punto, dos cromosomas son seleccionados de forma aleatoria a partir de la población intermedia; estos dos cromosomas actuarán como padres. Una posición del gen es escogida aleatoriamente y como resultado de este cruzamiento dos nuevos cromosomas son obtenidos.

En el cruzamiento uniforme, cada padre tiene la misma probabilidad de contribuir con sus genes para el único individuo resultante. Si un número generado aleatoriamente es más pequeño o igual que 0.5, entonces el gen será tomado del primer padre; en otro caso, será tomado del segundo padre.

Una vez que concluye el proceso de recombinación genética por medio del cruzamiento, si los cromosomas nuevos ya existen en la población, entonces es realizada una mutación para obtener cromosomas nuevos y diferentes.

La probabilidad de ocurrencia del cruzamiento estará definida por el usuario.

Operador de mutación

La implementación de este operador es muy simple. El operador tradicional de mutación se define de la siguiente forma: aleatoriamente se escoge un cromosoma, aleatoriamente también se selecciona el gen a mutar y se cambia su estado: 0 por 1 o 1 por 0, lo cual garantiza que la inclusión del clasificador se modifica en la combinación. Si el cromosoma resultante existe previamente, entonces se escoge otro punto de mutación y se repite el proceso. Si como resultado de explorar todos los puntos de mutación no se obtuviese ningún cromosoma nuevo, se selecciona otro cromosoma para mutar. La probabilidad de ocurrencia de la mutación estará definida por el usuario.

Operador de selección

Teniendo en cuenta las características del problema se utiliza el método de la ruleta, que no permite la selección de un individuo más de una vez.

En el método de la ruleta la probabilidad usada para cada cromosoma es calculada dividiendo el resultado de la función objetivo para el cromosoma entre la suma de la función objetivo de cada cromosoma en la población con tamaño P' . Esto se muestra en la siguiente fórmula:

$$p(C_i) = \frac{f(C_i)}{\sum_{i=1}^{P'} f(C_i)}$$

Para resumir se enuncian los pasos necesarios para el funcionamiento del AG. Cada iteración simple comienza con una población que tiene tamaño igual al número previamente especificado, esta población es generada usando un híbrido entre la generación aleatoria y el sembrado de individuos. Después de que una población intermedia es generada por el operador de selección, entonces los cromosomas nuevos son generados por el proceso de la recombinación, se agregarán para la población inicial y podrán estar o no en la nueva población.

La población será limpiada de cromosomas que probabilísticamente tomen los valores más pequeños en la función objetivo, hasta conservar el tamaño establecido (método de la ruleta).

El algoritmo para cuando sea cierto al menos una de las siguientes condiciones:

- El usuario especifica la parada cuando el algoritmo encuentra la primera combinación que satisface las condiciones y restricciones del problema.
- Se ha alcanzado el número de generaciones definidas por el usuario.

5. ANÁLISIS DE HIPERTENSIÓN ARTERIAL EN EDAD PEDIÁTRICA

El término de hipertensión arterial sistémica (HTA) es cada vez más común en nuestra sociedad y su identificación como factor de riesgo cardiovascular, sin embargo, no todo el mundo traslada esta preocupación a los niños. Las guías de la Sociedad Europea de Hipertensión (ESH) y de la Sociedad Europea de Cardiología (ESC) del tratamiento de la HTA, publicadas en 2003 y actualizadas en 2007, no incluyen, lamentablemente, ninguna sección dedicada a la HTA en niños y adolescentes [20].

La prevención de las enfermedades cardiovasculares no queda limitada a la edad adulta, sino que debe iniciarse en la edad pediátrica. La HTA es la mayor causa de morbimortalidad en muchos países, por sus consecuencias sobre el sistema cardiovascular y los accidentes cerebrovasculares. Se ha demostrado que la HTA en la infancia es un factor de riesgo independiente para la hipertensión en la edad adulta y está asociada con marcadores precoces de enfermedad cardiovascular como hipertrofia ventricular izquierda, espesor de la íntima-media, complianza arterial, aterosclerosis y disfunción diastólica. La prevalencia global de HTA en adultos es del 15-20%; mientras que, en niños con edades entre 4 y 15 años se estima en un 2%.

El diagnóstico de hipertensión en niños es complicado porque los valores normales y anormales de la presión sanguínea varían con la edad, el sexo y la talla, con un amplio rango descrito en tablas y por tanto, son difíciles de recordar. Se ha demostrado que la hipertensión en la infancia es un factor de riesgo independiente para la hipertensión en la edad adulta y está asociada con marcadores precoces de enfermedad cardiovascular. Considerando que la morbilidad y la mortalidad a largo plazo están asociadas a la hipertensión arterial, un componente importante para la salud de los niños y de los adolescentes es intervenir a tiempo [21].

5.1 Descripción de la muestra

En este estudio, la muestra estuvo constituida por un total de 680 niños supuestamente sanos entre 8 a 12 años de edad, de ambos sexos, pertenecientes a 4 escuelas primarias de la ciudad de Santa Clara. Los datos fueron suministrados por el proyecto PROCDEC de la Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba [22]. Se analiza un conjunto de 31 variables aleatorias empleadas en el diagnóstico de riesgo cardiovascular para predecir el riesgo de que un niño sea o no hipertenso. La tabla 3 muestra las características fundamentales de las variables aleatorias que son discretas y la tabla 4 muestra las características fundamentales de las variables aleatorias que son continuas.

Tabla 3: Variables predictoras discretas

Variables	Identificador	Valores	Porcentaje
Sexo	Sexo	Femenino	50.4
		Masculino	49.6
Color de la	ColorPiel	Blanca	81.5
		No blanca	18.5
Diagnóstico	newdiag	Sí	57.5
		No	42.5

Tabla 4: Variables predictoras continuas

Variable	Identificador	Mínimo	Máximo
Edad	EdadA	8	11
Peso actual (en kg)	PesoKg	20.5	75.1
Talla (en cm)	Talla	113	191
Circunferencia Cintura (en cm)	CCintura	37	182
Circunferencia Cadera (en cm)	CCadera	30	106
Índice Cintura Cadera	icc	0.58	2.5
TA Sistólica Miembro Inferior	TSMI	11	150
TA Diastólica Miembro Inferior	TDMI	20	90
TA Sistólica Miembro Superior	T1dBS	81	150
TA Diastólica Miembro Superior	T1dBD	6	99
TA Sistólica 5 min	T1d5S	73	152
TA Diastólica 5 min	T1d5D	46	95
TA Sistólica 10 min (antes 15 min)	T1d10S	11	149
TA Diastólica 10 min (antes 15 min)	T1d10D	45	94
TA 1 ^{er} día Sistólica Media	T1SMedia	77.33	146
TA 1 ^{er} día Diastólica Media	T1DMedia	43.33	88.67
TA 2 ^{do} día Sistólica Basal	T2dBS	82	150
TA 2 ^{do} día Diastólica Basal	T2dBD	17	94
TA 2 ^{do} día Sistólica P. Peso Sostenido	T2dPPSS	78	174
TA 2 ^{do} día Diastólica P. Peso Sostenido	T2dPPSD	48	116

Presión Arterial Media 2d	PAM2d	58	129.67
TA 3 ^{er} día Sistólica Basal	T3dBS	81	144
TA 3 ^{er} día Diastólica Basal	T3dBD	40	110
TA 3 ^{er} día Sistólica P. Peso Sostenido	T3dPPSS	83	160
TA 3 ^{er} día Diastólica P. Peso Sostenido	T3dPPSD	13	112
Presión Arterial Media 3d	PAM3d	53.33	120
Índice de masa corporal	imc	8.77	255.19

5.2 Modelos matemáticos

Se obtienen varios modelos de clasificación utilizando árboles de decisión, redes bayesianas, redes neuronales, regresión logística y otros incorporados en el software Weka (Waikato Environment for Knowledge Analysis) versión 3.7.5. Este software es libre, de código abierto, y tiene incorporado muchos métodos estadísticos y de Inteligencia Artificial.

En la tabla 5 se muestra la exactitud de estos modelos individuales obtenidos

Tabla 5: Exactitud de los clasificadores base

Clasificador	Exactitud
weka.classifiers.trees.FT (<i>Functional Tree structure</i>)	0.8831
weka.classifiers.lazy.LWL (<i>Locally weighted learning</i>)	0.7749
weka.classifiers.trees.RandomForest	0.8874
weka.classifiers.lazy.IBk	0.8225
weka.classifiers.functions.SPegasos	0.8874
weka.classifiers.trees.REPTree	0.8658
weka.classifiers.bayes.NaiveBayes	0.8225
weka.classifiers.functions.Logistic	0.9091
weka.classifiers.lazy.IBk	0.8182
weka.classifiers.trees.J48	0.8701
weka.classifiers.functions.MultilayerPerceptron	0.8701
weka.classifiers.trees.ADTree	0.9004
weka.classifiers.functions.SGD (<i>Stochastic Gradient Descent</i>)	0.9134
weka.classifiers.trees.RandomTree	0.8182
weka.classifiers.functions.SMO(<i>Sequential minimal optimization</i>)	0.9134
weka.classifiers.lazy.KStar	0.7013
weka.classifiers.functions.VotedPerceptron	0.5800
weka.classifiers.trees.LMT (<i>Logistic Model Trees</i>)	0.8961

Como se observa el mejor por ciento de clasificación se logra utilizando los clasificadores SGD y SMO, con una exactitud de 0.9134.

Luego combinando las 13 medidas de diversidad mencionadas en el epígrafe 3 (R, p, Q, D, DF, E, KW, k, DIF, GD, CFD, DFD, OD), utilizando el operador fuzzy mencionado anteriormente y teniendo como configuración del Algoritmo Genético: 50 generaciones, 0.25 probabilidad de mutación y 0.75 probabilidad de cruzamiento, se logran los resultados mostrados en la tabla 6.

En ella se muestra el cromosoma resultante del algoritmo genético y la exactitud del sistema multclasificador, como se observa este es un valor superior al mejor valor obtenido anteriormente.

Tabla 6 Resultados del Algoritmo Genético

C_x	Exactitud del sistema multclasificador
001001000100101100	0.9480

El cromosoma resultante corresponde a la combinación de los siguientes clasificadores:

weka.classifiers.trees.RandomForest,
weka.classifiers.trees.J48,

weka.classifiers.trees.REPTree,
weka.classifiers.functions.SGD,
weka.classifiers.functions.SMO,

weka.classifiers.lazy.KStar.

Nótese que con este cromosoma se mejora la clasificación individual en un 3%.

En la figura siguiente se muestra una comparación de los resultados de los clasificadores y el multclasificador en cuanto a la exactitud en la clasificación. Obsérvese que el sistema multclasificador supera a los clasificadores individuales.

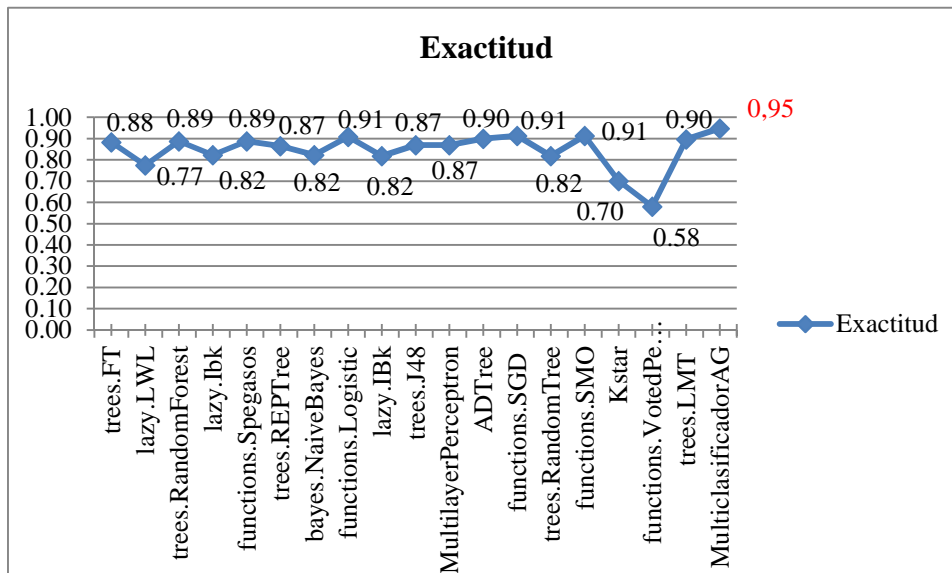


Figura 2: Comparación de los 19 modelos obtenidos en cuanto a la exactitud o por ciento de casos bien clasificados.

6. CONCLUSIONES

El presente trabajo muestra una técnica novedosa que emplea Algoritmos Genéticos para encontrar un buen conjunto de clasificadores diversos. La función objetivo del Algoritmo Genético involucra la exactitud del sistema multclasificador y los resultados de medidas de diversidad aplicadas a los clasificadores individuales del sistema.

Un caso de estudio de la base de HTA es usado para ejemplificar esta contribución. Se aplicaron en total 18 clasificadores base y sus resultados individuales no superan el 91%. Usando la propuesta del Algoritmo Genético con medidas de diversidad, se obtiene un multclasificador que logra mejorar en un 3% la clasificación anterior.

7. REFERENCIAS BIBLIOGRÁFICAS

- [1] Bonet, I. (2008), *Modelo para la clasificación de secuencias, en problemas de la bioinformática, usando técnicas de inteligencia artificial.*, in *Ciencias de la Computacion*. 2008, Universidad Central "Martha Abreu" de las Villas.: Santa Clara.
- [2] Mitchell, T.M. (1997) *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997: p. .
- [3] Quinlan, J.R. (1986) *Induction of decision trees*. *Machine Learning*, 1986: p. 1, 81-106.
- [4] Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*. San Mateo,CA, Morgan Kaufmann., 1993.
- [5] Chavez, M.D.C. (2008), *Modelos de redes bayesianas en el estudio de secuencias genómicas y otros problemas biomédicos.*, in *Ciencias de la Computacion*. 2008, Universidad Central "Martha Abreu" de las Villas.: Villa Clara.
- [6] Bello, R., et al. (2001) *Aplicaciones de la IA, Santa Clara*. 2001.
- [7] Wolpert, D. (1992) *Stacked generalization*. *Neural Networks*, 1992: p. 5, 241-259.
- [8] Salazar, S. (2005) *NEngine v1.0. Una Herramienta Software para Redes Neuronales Recurrentes*. . Universidad Central "Marta Abreu" de Las Villas., 2005.
- [9] Polikar, R. (2006), *Ensemble based systems in decision making*, in *IEEE Circuits and Systems Magazine*. . 2006.
- [10] Kuncheva, L.I. (2004) *Combining Pattern Classifiers: Methods and Algorithms*. New York, NY, Wiley Interscience., 2004.
- [11] Kuncheva, L.I. and C.J. Whitaker. (2003) *Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy*. . *Machine Learning*, 2003. **51**: p. 181-207.
- [12] Skalak, D.B. (1996) *The sources of increased accuracy for two proposed Boosting algorithms*. 1996.
- [13] Giacinto, G. and F. Roli. (2001) *Design of effective neural network ensembles for image classification purposes*. 2001.
- [14] Cunningham, P. and J. Carney. (2000) *Diversity versus Quality in Classification Ensembles Based on Feature Selection*, in *Machine Learning*. ECML , R. López de Mántaras and E. Plaza, Editors. 2000, Springer Berlin / Heidelberg., 2000: p. p. 109-116.
- [15] Kohavi, R. and W. D.H. (1996.) *Bias Plus Variance Decomposition for Zero-One Loss Functions in Machine Learning*. Proceedings of the Thirteenth International Conference. , 1996.
- [16] Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions*. John Wiley & Sons., 1981.
- [17] Hansen, L.K. and S. P. (1990.) *Neural Network Ensembles*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990. : p. 12: p. 993-1001.

- [18] Partridge, D. and K. W. (1997) *Software diversity: practical statistics for its measurement and exploitation*. Information and Software Technology, 1997: p. 39(10): p. 707-717.
- [19] Partridge, D. and W. Krzanowski. (1997) *Distinct failure diversity in multiversion software*. Res. Rep, 1997. **348**.
- [20] G. Mancia, et al. (2007) *Guidelines for the management of arterial hypertension: The Task Force for the Management of the Arterial Hypertension of the European Society of Hypertension (ESH) and the European Society of Cardiology (ESC)*. J Hypertens, 2007. **25**.
- [21] Ortigado, A. (2011) *Hipertensión arterial sistémica*. Tratado de Pediatría Extrahospitalaria. Madrid: Ergon, 2011: p. 455-62.
- [22] UCLV and U.D.O *Proyección del Centro de Desarrollo Electrónico hacia la Comunidad (PROCDEC)*