6th International Conference on Industrial Engineering and Industrial Management.
XVI Congreso de Ingeniería de Organización. Vigo, July 18-20, 2012

An Overnight Parcel Logistics Company's Capillary Distribution Network Design by Regression Analysis

Oscar Rioja San Martín[1], Joaquim Lloveras Maciá

Abstract

This article will explore and analyse how to design the capillary distribution network of an Overnight Parcel Logistics Company (hereby referred to as OPLC). The OPLC needs to design a capillary distribution network that is capable of collecting all the merchandise that their clients wish to deliver, and, at the same time, have the capacity to distribute all shipments to their destination. To reach this goal, the designers have to decide the type of vehicles (van, lorry or trailer) and the quantity of each type of vehicle the capillary distribution network needs to cover the area around each hub efficiently. This number of vehicles has to satisfy the delivery/collection quality requirements with the minimum cost to guarantee the maximum profit for the OPLC.

This research will demonstrate that regression analysis is a valid technique to help an OPLC in the decision making related to the distribution costs of the capillary network. The way the costs of each distribution areas are obtained for the design of the capillary network is by regression models. Most models related to logistics companies have been based on deterministic techniques so far.

Keywords: Capillary network, regression, distribution cost

1.1 Introduction

One of the main factors that affect an OPLC is the randomness of the goods it has to distribute. This randomness has a large effect on all sections of an OPLC, but the greatest is on the capillary distribution network. The OPLC's managers need to design a capillary distribution network that is capable of collecting all the merchandise that their clients wish to deliver, and, at the same time, this capillary distribution network must have the capacity to distribute all the

[1] Oscar Rioja San Martín (✉)
Technological Univesity of Catalonia
Barcelona, Spain
e-mail: errioxa007@hotmail.com

shipments to their respective destination.

To design this network, the most powerful tool the managers have is reliable and accurate information about the nature of the goods needed to be collected and delivered, and consequently, about the randomness of the merchandise. This information can be obtained through a model. For an external observer, a model of a system is an object that the observer can use to answer any question about the system he is interested in [Minsky, 1965]. So, once a model is created, it can be used to represent the behaviour of the real system.

In the design of a capillary distribution network, the information that is needed to create the most profitable network is the number of deliveries and collections for each area, and the weight of the different deliveries and collections. From the number of deliveries and collections, a manager will know the number of vehicles he needs for each area, and depending on the weight of the collections and the deliveries, the type of vehicles that are needed.

These two variables are fundamental for another critical aspect of the OPLC; the delivery and collection cost. In parcel logistics companies these costs are as follows: the parcel logistics company pays a specific amount to the vehicle in charge of the collection, one part of which is due to pick up the shipment, the other which depends on the weight of the shipment. The delivery cost is analogous; each delivery costs the parcel logistics company a specific amount which is divided into two parts; a fixed amount for serving the shipment, and another one based on the weight of the shipment.

For the design of a capillary distribution network it is important to consider that the delivery and collection costs for the OPLC are the income of the drivers of the network. Due to the particularity of the delivery cost (and collection cost: they are analogous), the relationship between the cost and the number of deliveries and their weight cannot be predicted exactly. The model created to predict the behaviour of these costs has to be able to deal with their specific characteristics of uncertainty. Because of the uncertainty and random components, a stochastic model is required [Guasch et al, 2002].

In the literature related to logistics companies, regardless of the nature of the goods they distribute or the delivery time they offer to their clients, most of the models are created by deterministic methods. Mostly, all the aspects of logistics companies have been modelled by deterministic models, e.g. numbers of vehicles, route selection, numbers of hubs and their locations and truck loading times and flows (cross docking).

Stochastic methods have not been used extensively in researches related to logistics companies. In most cases, these stochastic models have been used as a complement of the deterministic models, adding certain stochastic characteristics. An example of this is the Vehicle Routing Problem (VRP), which was renamed the Stochastic Vehicle Routing Problem (SVRP) when

certain stochastic parameters were included. VRP models are based on a number of locations to be visited, and certain characteristics, such as the weight of the deliveries and the collections, are known. However, it is more common that these characteristics are not previously known. Parameters such as costumer demand, the travel time between clients, and even locations to visit are stochastic in nature. For this reason, the VRP models were renamed as SVRP - the stochastic vehicle routing problem. The stochastic part of SVRP models are based on Markovian stochastic processes [Gendreau et al, 1996].

Other investigations, however, assign vehicles to certain distribution areas instead of working with a determined route of clients and visits for every vehicle. These researches are focused on the familiarity of the distribution vehicle drivers with their distribution area. Here, the productivity of the distribution vehicles is analysed when the familiarity of a driver with its distribution area increases. With increased familiarity, driver performance improves due to ease in finding addresses and locations, and efficiency in organising daily routes. In this way, their ability to make deliveries and collections increases, and therefore, their productivity [Zhong, Hall et al, 2004].

This line of investigation will be followed in this paper. The distribution network will be divided into distribution areas based on the drivers' income in each area, instead of assigning a determined number of clients or a determined route. These distribution areas will be based on postcodes and the drivers' income in each area will be predicted. If this income is not high enough, a driver can have more than one postcode. If the income is too high, more than one vehicle can share the same postcode.

The aim of assigning each vehicle to specific postcode areas is to guarantee a minimum income, thus to ensure the continuity, of its driver, because this continuity maximises driver familiarity within their distribution area. With increased familiarity, driver performance improves due to ease in finding addresses and locations as well as efficiency in organising daily routes. Their capacity to make deliveries and collections increases, and therefore, so does their productivity. The way the drivers' incomes model of each distribution areas is obtained is by regression analysis. The regression analysis is the most used technique nowadays and because of its multiple uses, regression applications are numerous and there are in almost any field [Montgomery et al, 2004]. In the reviewed literature, cases of capillary distribution networks modelled by regression analysis have not been found. This research will demonstrate that regression analysis is a valid technique to develop a model which helps a parcel logistics company's manager in the decision making related to the distribution costs of the capillary network, and therefore, in predicting the incomes of the various vehicle drivers.

1.2 Case Study

The case study for this research is a hub of an OPLC located in Barberá del Vallés, in the metropolitan area of Barcelona, Spain. This hub has several towns within its influence zone, and in this paper, one town called Sant Cugat del Vallés, composed of seven postcodes, will be studied. Sant Cugat del Vallés is located 11.3 km from the OPLC's hub, and they are connected by the A-7 highway. The vehicle income of each one of these postcodes will be predicted to analyze if a vehicle can be assigned to each one.

The vehicles that work in this particular OPLC are in a production regime. Each OPLC has its distribution vehicles in the regime most beneficial for the company, but in most cases a production regime is selected: the more deliveries and collections they make, the more they earn. The vehicle drivers do not have fixed incomes and their incomes depend on the number of deliveries and collections they do.

The income of each delivery and collection is tabulated in different strata based on its weight. A fixed price is assigned to the drivers of the vehicles due to the delivery of each shipment or each collection done. Another variable price is added to the fixed one depending on the weight of each shipment they deliver or collect. Both, the fixed and variable part, depend on the weight of every delivered shipment or collection. The income for a delivery does not have to be the same as for a collection, and the variation among the prices of the different strata does not have to be linear. Every OPLC regulates these prices in the way it believes best suits its interests.

1.3 Methodology

As explained in the introduction, the distribution vehicles' incomes per postcode area will be modelled by regression analysis. Regression analysis is a statistical methodology that uses a relationship between two or more quantitative variables, in such a way that a response or output variable is related to one or more of these quantitative variables [Kutner et al, 2004]. So, in order to perform a regression analysis, first it is necessary to know which output variable will be modelled, i.e. which is the endogenous variable. Secondly; which other variables are related with the one to be modelled, i.e. which are the exogenous variables. And finally, it is necessary to know the relationship between the endogenous variable and the exogenous variables: how the endogenous variable can be estimated by the exogenous ones.

1.3.1 Variable Identification

The income of one vehicle per postcode area can be divided into the delivery income and collection income. Therefore, one model will be developed to estimate the delivery income, and a different model will be developed to estimate the collection income; these incomes being the

endogenous variable of their respective models. The delivery income will be estimated by the number of deliveries per day and the total delivered weight per day, considered as the exogenous variables. In the regression analysis these two variables will be raised to the second power; four endogenous variables in total. The collection income is analogous: the number of collections per day and the total collected weight per day, *and* these variables raised to the second power, are the exogenous variables. As the exogenous variables are expressed per day, the delivery and collection income will also be estimated per day.

Once all the variables that are related with the distribution incomes have been identified, the next step is the sampling of them. Data collection is one of the most laborious stages in a model construction, yet it is of utmost importance for achieving an efficient model. Any model is only as good as the data on which it is based [Vincent, 1998]. For this research, the sampling will be an observational study, where the system is not affected by the sampling, obtaining all the data from the OPLC database. For each one of the six variables, endogenous and exogenous, 84 days of data will be obtained, providing 84 observations per variable in each one of the postcode areas that are to be analysed. Afterwards, the independence of the 84 observations of each one of the variables has to be checked to ensure that there is not any correlation between them. To do this, two heuristic tests will be undertaken: correlation graphs and dispersion diagrams [Guasch et al, 2003].

Once the independence of the observations of each variable is checked, the observations of each variable will be divided into two groups. The first, called the calibration sample, will be used for the regression analysis and to create the model which will estimate the delivery and collection costs. The second group, the prediction sample, will be used to check the validity and the prediction capacity of the model [Snee, 1977]. This separation will be done by random assignment.

1.3.2. Regression Model Elaboration

When the endogenous and exogenous variables have been identified and their sampling has been completed and checked, the next step is to analyze the relationship between the endogenous variable and the exogenous variables: the regression model has to be elaborated or adjusted. For that, the regression parameters for each one of the endogenous variables have to be calculated. The regression model adjustment will be developed by maximum likelihood. The maximum likelihood method is an interactive procedure. With this method, assertions or statements about the unknown parameters θ of a probability distribution or a function can be made from the observed data of a sample [Hocking, 2003]. For a sample composed of independent observations y_i the equation that defines maximum likelihood is:

$$L(\theta) = f(y;\theta) = \prod_{j=1}^{n} f(y_i;\theta) \tag{1.1}$$

This expression is known as the maximum likelihood function. The assumption that *f* is known except for the uncertainty of θ, reduces the problem of making assertions about the plausible values of θ, since y_i values are given [Davison, 2003]. The maximum likelihood function quantifies the possibility that θ generates the observed sample values. The higher the value of the maximum likelihood, the more likely that θ generates the observed sample values. In the case of regression analysis, the θ parameters will be the unknown regression parameters of each of the exogenous variables. Using a spreadsheet, the value of the regression parameters that maximize the maximum likelihood function value will be determined.

The main advantage of the use of this iterative procedure is that without very complicated modifications it enables the adjustment of the regression models with different characteristics. For example, the least squares method assumes that the residuals are distributed according to a normal probability distribution function. Using the maximum likelihood method, you can adjust the model under this assumption, but it also allows you to adjust it assuming that residuals are distributed with any other probability distribution functions. It is possible to check if these models with residuals distributed with non-normal probability distributions increase the accuracy of the regression analysis or not, by comparing the maximum likelihood values. Using maximum likelihood, the model can also be adjusted under the principles of generalised linear models and compared with the adjusted models previously carried out, thus checking if a non-linear regression analysis better fits the data than a linear one, by comparing again the maximum likelihood values.

For this research, three types of regression models will be compared, which have been adjusted under these three different assumptions:

- Ordinary least squares.
- The regression analysis residuals' probability distribution function is double exponential or logistic
- Under the criterion of generalized linear models

1.3.3 Model Selection

Once the procedure with which the regression model will be developed is known, and before carrying it out, it is necessary to consider the procedure that will determine which exogenous variables really influence the behaviour of the endogenous variable in a substantial way. That is, the model has to respect the parsimony principle: the best model is the simplest one that explains the observed facts [Navidi, 2006].

The maximum likelihood method offers the possibility of determining to what extent the addition of a new variable to the model provides a better fit to the data, using the likelihood ratio test. The likelihood ratio test will determine whether or not the inclusion of certain variables increases the accuracy of the regression analysis when adjusting the model. This test compares the maximum likelihood value obtained with $k+1$ variables with that obtained for k variables, and if the difference is not higher than a freedom degree of a X^2 probability distribution function for the selected significance level, the new $k+1$ variable will not be included in the model.

In short, the maximum likelihood method has been chosen for its flexibility in adjusting the regression models under different assumptions and the ability to compare them by observing the maximum likelihood values. The maximum likelihood method also shows whether adding new variables to the model provides a better fit to the data, to fulfil the parsimony principle

1.3.4 Model Validation

After the adjustment of the different models and identifying which one fits better to the data, the next stage in the development of the model is its validation. For this, the data from the prediction sample will be used instead of the data from the calibration sample, which was used for the adjustment. The objective of model validation is to verify that the model meets the requirements according to objective and subjective criteria under which it was conceived [Johansson, 1993].

The validation procedure consists of introducing the exogenous variable data from the prediction sample into the model, and comparing the outputs provided by it with the correspondent endogenous variable data of the prediction sample. As the observations of the prediction sample are independent, classical statistical methods can be applied. For this research, Welch test or paired-t test will be used [Law and Kelton, 1991].

1.4 Conclusions

In this research the hypotheses made in the introduction of this work have been solved successfully: regression analysis is a valid technique to help a parcel logistics company in the decision making related to the distribution costs of the capillary network, and therefore, in predicting the incomes of the different vehicle drivers. The delivery and collection incomes of seven postcode areas have been correctly modelled, and it is possible to ascertain if each one of these postcode areas can be assigned to a vehicle, or on the contrary, more than one postcode should be assigned to a vehicle to guarantee a minimum income for the vehicle driver. To estimate these incomes the number of deliveries and collections per day and the total delivered and collected weight per day needs to be known.

For the cases studied in this research, the linear regression models have had better results than the non-linear ones. The models elaborated under the assumption of the ordinary least square and the residuals distributed by double exponential and logistic probability distribution functions have had better results than the ones elaborated under the assumption of the exogenous variables' transformation. The first models have provided a better fit to the data of the samples, their maximum likelihood values being higher than the ones of the models made under the generalized linear models criteria.

Of the seven delivery income models created, four models adjusted under the assumption of the ordinary least square have obtained the highest maximum likelihood values. For the other three models, the highest maximum likelihood values were obtained under the assumption that the residuals are distributed according to a logistic distribution function. In the case of the collection income models, five out of seven models adjusted under the assumption of the ordinary least square provided the highest maximum likelihood values. Of the other two, one was adjusted supposing that the residuals are distributed according to a logistic distribution function, and the other, according to a double exponential probability distribution function. In the cases where the models have been adjusted under the supposition of the ordinary least square and the maximum likelihood value has not been the highest, it has been proved that these ordinary least square models can still successfully predict the delivery or collection income. So, the ordinary least squares method has been proved as an appropriate regression technique to create the models with which the distribution costs can be estimated.

1.5 References

Davison, A.C. (2003) Statistical models. Cambridge University Press

Guasch, A., Piera, M. A, Casanovas, J., Figueras, J., (2002) Modelado y simulación. Aplicación a procesos logísticos de fabricación y servicios. Edicions UPC, Barcelona

Gendreau, M., Laporte, g., Segini, R. (1996) Stochastic vehicle routing. European Journal Of Operational Research

Hocking, R. R. (2003) Methods and applications of linear models. John Willey and Sons, New Jersey

Johansson, R., (1993) System modelling & identification. Prentice Hall, New Jersey

Kutner, M. H., Nachtsheim C. J., Neter, J. (2004) Applied linear regression models. Mc Graw Hil, New York

Law, A. M, Kelton, D. W. (1991) Simulation modelling and analysis. Mc Graw Hill, New York

Minsky, M., (1965) Matter, Mind and Models. Spartan Books, Washington D.C.

Montgomery, D. et al (2002) Introducción al análisis de regresión lineal. Compañía editorial Continental, México D.F.

Snee, R.D. (1997) Validation of regression models: Methods and examples. Technometrics

NavidiI, W. (2006) Estadística para ingenieros y científicos. Mac Graw Hill, México D.F.

Vincent, S, (1998) Input data Analysis. Handbook of simulation. Ed. Banks, J., John Wiley & Sons, New Jersey

Zhong, J., Hall, R.W., Dessouky, M., (2004) Territory planning and vehicle dispatching. University of Southern California, Los Angeles