CIT2016 – XII Congreso de Ingeniería del Transporte
València, Universitat Politècnica de València, 2016.
DOI: http://dx.doi.org/10.4995/CIT2016.2016.4058

XII Conference on Transport Engineering, CIT 2016, 7-9 June 2016, Valencia, Spain

INFORMATION RELATED TO POSTAL FLOWS AND BIG DATA ANALYSIS POTENTIAL. THE CASE OF SPAIN

Oscar Martínez-Alvaro[a]* , Angela Nuñez-González[b]

[a] *Senior researcher. Transyt (UPM). ETS Caminos. Prof. Aranguren s/n. Madrid. 28040. Spain.*
[B] *Junior researcher. Transyt (UPM). ETS Caminos. Prof. Aranguren s/n. Madrid. 28040. Spain.*

Abstract

National Post Offices manage huge volumes of letters and parcels. Data associated to these flows are growing fast, with a great variety related to the diversity of postal products. The research described in this paper has classified all information flows of Correos, the Spanish National Post Office. In spite of the complexity of the current postal service portfolio, only four categories of matrices allow the classification of all postal information flows. Thanks to the migration towards new products, analyses with simple techniques will provide more and better information in the future, due to the structured nature of existing databases.

Keywords: Post Office, Postal service, Big data, Origin destination matrix, Logistics, Information flow, Transportation flow.

1. Introduction

Big data is a buzzword and in recent times many successful companies have something to do with big data. Recent research by Özköse *et al*. [1] shows that firms with the ability to store, process and put into service a vast amount of data are deemed to have a clear advantage in the market, with companies like Google, Amazon, Facebook, YouTube and eBay being the most conspicuous examples of this assertion.

But this trend is not limited to big companies doing business in a specific sector. Postal Operators manage huge volumes of shipments every year and there is an increase in associated data due to a growing network of sensors that collect information and communicate with other devices, computers and systems. Until now, an important field of activity has been related to operation optimization, where important improvements in efficiency could be met. For example, according to Hassall *et al*. [2], Australia Post developed a Transport Information System that recorded the activity of vehicles, the products carried and the network points services on a "transport duty" basis. Thanks to this system, higher levels of productivity through better planning allowed a 14% reduction of fleet vehicles over a ten year period, in spite of a growth of mail and parcel volumes.

But changes in the postal sector are transforming it beyond what could be expected a few years ago, as stated by Norman [3]. On the one hand, new technologies are now posing an important threat to postal companies, due to the migration from physical mails to electronic mails. On the other, new business models have arisen and parcels are a fast growing market segment. Therefore, postal companies are facing a critical period of transition and most of them are embarked on sophisticated programmes, such as the one described by Zeiler [4] about the U.S. Postal Service, who has launched a project to build an *IoPT* ("Internet of the Postal Things") expecting to reduce its marginal costs. But , according to Hu [5], the applications of big data technologies could be much more than a tool of cost reduction and be a key factor in the development and transformation of the postal sector, citing the example of China Post, which in 2010 established a data analysis team in charge of data analysis and integration, to support the marketing department.

On the other hand, the potential use of postal data beyond the internal use for Postal Operators has been recognised for long. More than 20 years ago Mori [6] analysed interrelationships among regions based on freight, postal origin-destination flows, and telecommunications in Japan, calibrating a mathematical model. More recently, according to the U.S. Postal Service [7], the postal industry has begun to experiment with Big Data analyses in several projects for many different purposes. For instance, in 2014, DHL was developing a tool to analyse correlations between different data sources such as weather, flu epidemics and Google trends to predict parcel volumes and determine staff and vehicle requirements.

And the use of postal data that began as a management tool, can benefit other entities. For instance, Banarjee [8] shows that Postal Operators began with geomodelling long ago for internal purposes, and Zhang [9] expands the opportunity for services to third parties that would welcome further geoinformation, while Nagabhushan *et al.* [10] deem this phenomenon to be specifically beneficial in developing countries where addresses are incomplete or simply approximate. And applications could be found in fields as remote as public health research according to Dankar *et al*. [11].

Adding to all this, according to Houck [12], there are attempts by some Postal Operators to enrich their routine daily tasks including collecting data about traffic (speed and delay), pollution (concentration of pollutants), road conditions (pavement condition), etc. Certainly, these data could be of great interest, say, to governments in order to adapt their road maintenance priorities or to identify sources of pollution, but would require an additional investment in the form of sensors and specific devices.

In the midst of this complex panorama, the aim of this paper is just intermediate. First, it is related to the use of postal information that could be of interest for third parties, not only for internal purposes of Postal Operators. Second, it is focused on existing information, not having to create new databases, to design new processes or to implement new recording devices.

2. TYPICAL POSTAL DISTRIBUTION STRUCTURE

All National Postal Operators follow a similar pattern, and the case of *Correos*, the Spanish Postal Operator, may be taken as typical. Operations of the postal service can be classified in abstract into the following groups of processes:

- Collection & admission: operations of collecting or receiving postal items at access points of the postal network (mailboxes, post offices, etc.).

- Classification: operations aimed at grouping all received objects according to their destination.
- Transportation: the transfer of postal items from the place where they have been classified to the addressee.

In fact, the real word is much more complex. To begin with, the classification is carried out within a structure of hubs, where postal objects may be classified in one or several steps, depending on the geographical origin and destination of the postal object, as well as on the hierarchical structure of hubs. In the most typical case, a postal object after entering the postal system is sent to the closest postal hub where it is pre-classified and aggregated to other objects that are sent to another hub that has the destination address within its area of influence. In this later hub, objects are classified again and sent to the corresponding local post office. In these local offices, postal objects are divided into small batches that are distributed by postmen who follow predefined delivery routes.

This process may vary according to product lines. Nowadays, any postal office, such as *Correos*, offers a wide range of products. The classical letter may be sent as ordinary or registered mail. And companies may use the Post Office for sending bags of letters as internal mail between their premises. And there are also parcels (larger units), telegrams (sent electronically, but delivered physically), money remittances (subject to the strictest identity controls), etc. And most of them may be either ordinary or urgent. And so forth.

As a result, postal networks have typically a hybrid hub-and-spoke structure, with a high degree of complexity in order to provide efficient delivery processes for such a complex array of products. This complexity has led to many analyses oriented towards the optimization of such systems, such as the one carried out by Lee [13] for the case of Korea Post.

3. A THEORETICAL SYSTEMATIZATION OF FLOW ANALYSIS

To better understand the flow data availability, it is useful to follow the path of a hypothetical “postal object”. The first step of a “postal objects” (say, letters o parcels) occur when senders place their shipment in admission points, which may be a mailbox, an office, or other admission point (such as massive reception offices, where vanloads of letters, which may be classified by destination or not, are received from big clients, such as utilities). In all admission points (except post-boxes), the total number of postal objects received is registered in databases. Only for specific products the sender and the addressee are recorded at admission points.

All objects are transferred from their admission point to the closest hub classification centre. The process control at each classification centre (C_i) provides the sum of all inflows (I_i), but not its breakdown:

$$I_i = K_i + B_i + O_i + X_i \qquad (1)$$

Where,

$K_i = \sum k_{ij}$ = shipments whose origin is another country.
$B_i = \sum b_{ij}$ = shipments whose origin is a post-box.
$O_i = \sum o_{ij}$ = shipments whose origin is an office.
$X_i = \sum x_{ij}$ = shipments whose origin is another admission point.

After the pre-classification in the first Centre (C_i), shipments are sent to a second classification Centre (C_j). The volume of shipments from a Centre to every other C_n (origin Centre C_i itself included) is recorded: the total outbound flow O_i is composed of known volumes Ω_{ij} whose origin is Centre i and destinations j are other Centres. This is shown in equation (2):

$$O_i = \Omega_{i1} + \Omega_{i2} + ... + \Omega_{in} \qquad (2)$$

Where,

n = total number of classification centres

And the balance of inbound and outbound flows is given by the condition (3), developed by equation (4):

$$I_i = O_i \tag{3}$$

$$K_i + B_i + O_i + X_i = \Omega_{i1} + \Omega_{i2} + ... + \Omega_{in} \tag{4}$$

Within the destination Centres (C_j), the information about the destination of each shipment depends on the kind of product. The outflow Δ_j is characterized by equation (5):

$$\Delta_j = K_j + P_j + D_j + A_j \tag{5}$$

Where,

$K_j = \sum k_i$ = Shipments with destination in another country.
$P_j = \sum p_i$ = Shipments with information on postal code of addressee.
$D_j = \sum d_i$ = Shipments information on distribution zone of addressee.
$A_j = \sum a_i$ = Shipments with full information of addressee.

The balance of inbound and outbound flows in the centre C_j is given by equation (6), developed by (7):

$$\Omega_{1j} + \Omega_{2j} + ... + \Omega_{ij} = \Delta_j \tag{6}$$
$$\Omega_{1j} + \Omega_{2j} +... + \Omega_{nj} = K_j + P_j + D_j + A_j \tag{7}$$

That is, there is some potential of systematisation in order to describe some attributes of postal flows. Although only a small portion of postal products have detailed information on sender and addressee, there are huge volumes of information related to origins, destinations and intermediate nodes (such as classification hubs). But current information systems provide less information than could be expected, since operational and economic priorities, amidst highly competitive strains, make it impractical to collect all potential information. And mathematical algorithms cannot be of great help due to the highly undetermined nature of the equation systems shown above.

4. A PRACTICAL SYSTEMATIZATION OF INFORMATION AVAILABLE

To complicate things further, the usual, commercial classification of postal products leads to not less than several dozens of types. This wide range of products makes it very complex the logistics associated and, as a consequence, there is a vast array of processes of data collection and storage.

But the nature of information associated to many products is quite similar. Consequently, the whole set of products may be categorized into four groups, each homogeneous from the point of view of information associated. Table 1 presents this simple classification, based upon nature of information flows, not upon the commercial orientation or the pricing policy.

Table 1. Groups and product types according to information associated.

Group	Products
A	Ordinary mail, brochures, magazines, books, and other special products
B	Express mail, both national and international
C	Registered mail, both national and international

D	Parcels, company bags, international express, fax, money remittances, telegrams, and other value-added products

Source: Own elaboration.

Shipments of **group A** have high quality data in terms of total shipments by origins or destinations. Additionally, outflows and inflows among classification centres are well characterized. But no complete origin and destination information by shipment exists and it is not feasible to collect it. In this first group of products, reliable data are available about products entering each admission point, except for the mailboxes whose only known data are the total volume by influence area. There is also information available about shipments delivered at each postal code or distribution area, with the same limitations as above (totals are known, but not their breakdowns). On the other hand, shipments received in a classification centre originated abroad and shipments with destination abroad are known, but only as totals.

Somehow surprisingly at first sight, products of **group B** have a low quality of information when compared to group A. Operational requirements are the priority. Reliable data are available about number of products originated at each admission point of the postal network, except for the mailboxes, of which there are not any available data. Additionally, reliable information is available about import flows classified in all hubs and shipments with an international destination. No breakdown is available on a lesser detail.

When it comes to **group C**, its information quality could be in theory almost perfect, but economic limitations make it currently slightly similar or better than group A, depending on the type of product. Reliable data are available about products entering at each admission point of the postal network. Likewise, data are available about total number of products delivered at each postal code, but there is no deeper breakdown. Once again, reliable information is available about import flows classified at each centre and export flows at each admission point.

Group D has the highest information quality: reliable information about origin and destination is available, thanks to a complete traceability of each shipment. That is, all origin-destination pairs are known to postal address detail.

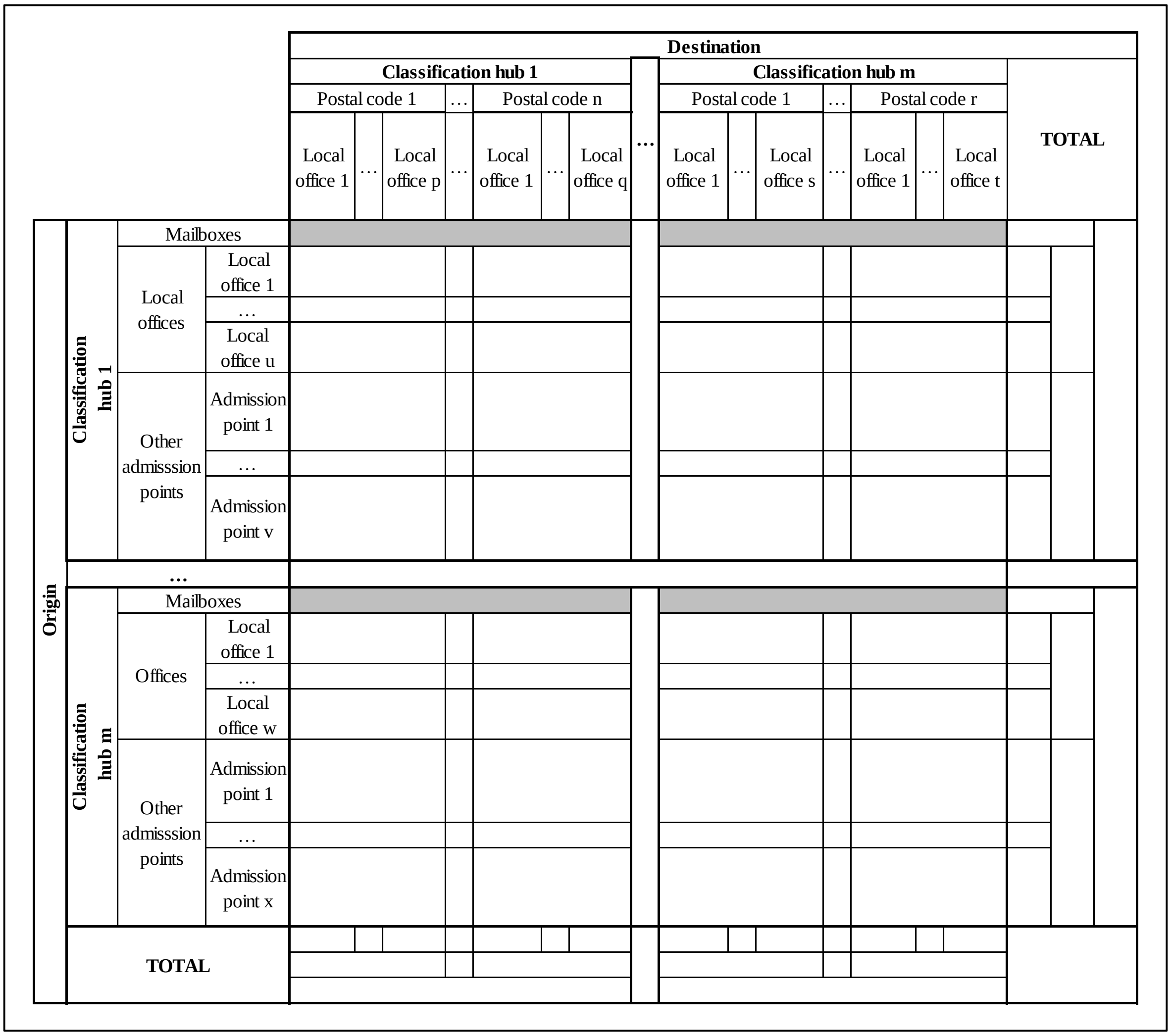
5. MATRIX STRUCTURE OF INFORMATION AVAILABLE

The structure of information described above is complex, but in fact does not reflect the real picture: a whole description of all details will require several dozen pages. Nevertheless, the primary groups A to D pave the way for a full, systematic classification of all flows, even including international ones.

The tool used for this purpose is the origin-destination matrix, so dear to transport planners. But in this case the zoning system must be defined on an *ad hoc* basis, since actual zoning is different for origin and destination. With this approach, every group A to D has a basic matrix, with a variant for international flows. The basic matrix shows, in some cases, some different aggregations depending on the product: for some products there is information for individual cells or sums of individual cells, while for others the information is only for total values of rows (origins) or columns (destinations).

Fig 1 shows the basic matrix for ordinary mail within Group A, as a conspicuous example. This matrix can be obtained from current information systems. Besides, a higher aggregation (flows between classification hubs) is provided by a different source of information, this giving a confirmation of information consistency.

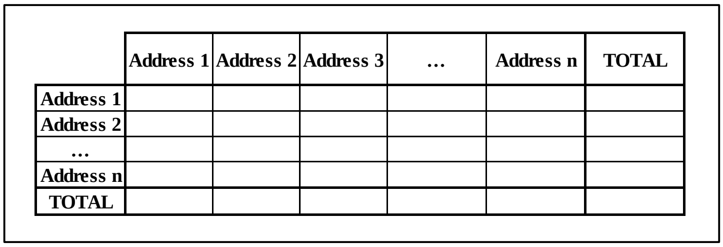
DOI: http://dx.doi.org/10.4995/CIT2016.2016.4058



				Destination															
				Classification hub 1							…	Classification hub m							TOTAL
				Postal code 1			…	Postal code n				Postal code 1			…	Postal code r			
				Local office 1	…	Local office p	…	Local office 1	…	Local office q		Local office 1	…	Local office s	…	Local office 1	…	Local office t	
Origin	Classification hub 1	Mailboxes		▨	▨	▨	▨	▨	▨	▨		▨	▨	▨	▨	▨	▨	▨	
		Local offices	Local office 1																
			…																
			Local office u																
		Other admisssion points	Admission point 1																
			…																
			Admission point v																
	…																		
	Classification hub m	Mailboxes		▨	▨	▨	▨	▨	▨	▨		▨	▨	▨	▨	▨	▨	▨	
		Offices	Local office 1																
			…																
			Local office w																
		Other admisssion points	Admission point 1																
			…																
			Admission point x																
	TOTAL																		

Fig. 1 - Flow matrix for Group A. Ordinary mail.
Legend: Shadowed cells not available. Source: Own elaboration.

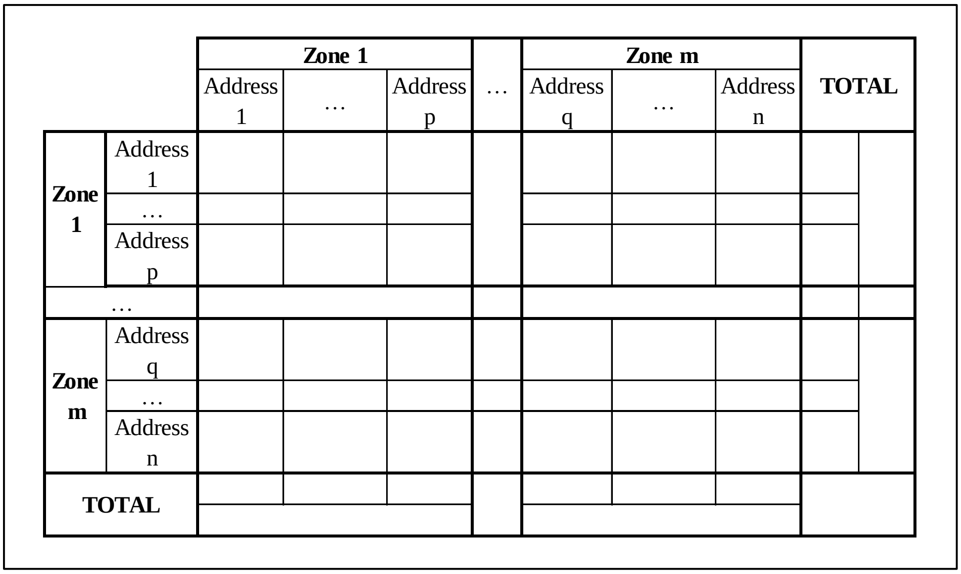
The extreme example of disaggregation is Group C: its flow information can be as detailed as on address basis, as shown in fig 2. But the very detailed nature of this information makes it impractical. The advantage is that it can be consolidated into zones with great flexibility depending on the objective of the analysis (streets, neighbourhoods, postal districts, cities, etc.), as shown in fig 3.



	Address 1	Address 2	Address 3	…	Address n	TOTAL
Address 1						
Address 2						
…						
Address n						
TOTAL						

Fig. 2 - Flow matrix for Group D. No aggregation.

Source: Own elaboration.



		Zone 1			…	**Zone m**			**TOTAL**
		Address 1	…	Address p		Address q	…	Address n	
Zone 1	Address 1								
	…								
	Address p								
…									
Zone m	Address q								
	…								
	Address n								
TOTAL									

Fig. 3 - Flow matrix for Group D aggregated on an ad hoc zoning basis.
Source: Own elaboration.

6. VOLUME BY INFORMATION GROUP

New technologies are impacting the postal market in two different ways. On the one hand, the supply chain is taking advantage of technology innovation and increased quality and quantity of information, as mentioned above. On the other, the demand side is changing fast: e-commerce is increasing the number of parcels sent, but e-substitution is on the path of eliminating traditional letters due to e-mails and e-chats, as shown by the aforementioned Norman [3].

Group A is the most numerous by far and B is the smallest, also by far. But the trend is not related to current volume: while volumes of Groups A and B are decreasing fast (between 5% and 10% per year in the last few years), Groups C and D show an upward trend. On the other hand, shipments in group D have the highest information quantity, followed by Group C. All this is summarised in table 2, where it is easy to see that products with better information are growing, while the opposite is true for products with lesser information.

Table 2. Volume and information quantity of information groups.

Group	Current volume	Volume trend	Information quantity per shipment
A	✚✚✚✚	▼	✚✚
B	✚	▼	✚
C	✚✚✚	▲	✚✚✚
D	✚✚	▲	✚✚✚✚

Note: Number of symbols of current volume is proportional to log scale of actual figures. Number of symbols of volume trend and information quality is proportional to linear scale of actual figures. Source: Own elaboration.

7. CONCLUSIONS

Currently there is an important volume of data about postal flows. Even without any further improvements in data collection, important volumes of information are available both about origins and about destinations of shipments, with different detail levels depending on the postal product type.

But the good news is that the types of postal products with an increasing trend are these whose information is of the highest quality. And this means that not only the information will increase, but that the nature of information will change: current information is reasonably good about areas of origin or destination of shipments, but the trend is towards precise information about addresses of origin and destination.

This will change the nature of the possibilities for exploiting postal data, well beyond their mere volume. The structured nature of the postal data collected by routine operations does not require any additional financial investment to the one need for operational purposes, and its analysis will be much simpler than the one needed with another type of data such as tweets, Google searches, position of mobile phones and the like.

All these are very optimistic news on the future of postal data and their potential use for many purposes. The only caveat is related to privacy, which has to be considered very carefully: probably new privacy policies adapted to the new scenario have to be properly defined and implemented.

8. ACKNOWLEDGEMENTS

This article is based upon research carried out within the "*Cátedra Correos*" (Correos Chair) funded by *Correos* (Spanish Postal Operator) by means of an agreement with the UPM. The analysis and interpretation of data, its conclusions, the writing of the report, and the decision to submit the article for publication are the sole responsibility of the authors.

9. REFERENCES

[1] Özköse H, Ari E. S, Gencer C.. Yesterday, Today and Tomorrow of Big Data. *Procedia-Social and Behavioral Sciences*. 2015. 195, p. 1042-1050.
[2] Hassall K., Sluyter P, Scott D. The Development of a Transport Meta-Language to Achieve Urban Freight Efficiencies: A Case Study of the Development and Application of the Australian Postal Corporation's 'Transport Information System'. *Procedia-Social and Behavioral Sciences*. 2012. 39, p. 282-292.
[3] Norman, H. Executive forum: six experts share their views. *Postal technology international.* Awards Special Issue. 2014. Annual showcase (1), p. 18-27.
[4] Zeiler, K. The Internet of things. *Postal technology international*. September 2014 (1), p. 23.
[5] Hu X, Jin Y, Wang F. Research of Postal Data mining system based on big data. *3rd International Conference on Mechatronics, Robotics and Automation*. 2015.
[6] Mori S. A structure analysis of interrelationships among regions based on freight, postal OD flows, and telecommunications in Japan. *Telematics and Informatics*. 1994. 11(3), p. 237-253.
[7] U.S. Postal Service. Office of Inspector General *International Postal Big Data: Discussion Forum Recap*. 2014. Report Number: RARC-IB-14-002.
[8] Banerjee S, Gelfand A. E, Polasek W. Geostatistical modelling for spatial interaction data with application to postal service performance. *Journal of statistical planning and inference*. 2000. 90(1), p. 87-105.
[9] Zhang M, Meng L. An iterative road-matching approach for the integration of postal data. *Computers, Environment and Urban Systems*. 2007. 31(5), p. 597-615.
[10] Nagabushan P, Angadi S. A., Anami B. S. A soft computing model for mapping incomplete/approximate postal addresses to mail delivery points. *Applied Soft Computing*. 2009. 9(2), p. 806-816.
[11] Dankar F. K., El Emam K, & Matwins S. Efficient Private Information Retrieval for Geographical Aggregation. *Procedia Computer Science*. 2014. 37, p. 497-502.
[12] Houck, K. A. (2014). Big Data. *Postal technology international*. September 2014(1), p. 20-28
[13] Lee J. H, Moon I. A hybrid hub-and-spoke postal logistics network with realistic restrictions: A case study of Korea Post. *Expert systems with applications*. 2014. 41(11), p. 5509-5519.