Human elicited features in retail site analytics

Hui-Jia Yee, Choo-Yee Ting, Chiung Ching Ho

Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia.

huiijiia@gmail.com, cyting@staff.mmu.edu.my, ccho@mmu.edu.my

Abstract. Location selection is indispensable for a company or industry to survive for a long term. A strategically positioned retail location can increase the profitability and draw more customers for a company. Conventionally, a good location decision is associated with the relevant and significant location factors. However, integrating human subjective opinion as part of feature engineering process can be a challenge; there is no guarantee that these features can be optimum. In this light, this paper aims to investigate the impact of additional human elicitation features on the retail site selection model. This research focuses on retail site analytics to predict the sale of a telecommunication company in Malaysia. Apart from features such as geographical information, demographics and economics, this paper also includes the features determined by domain experts to investigate whether the human elicited features could improve the accuracy of sales estimation given a specific location. The findings of current work show that the additional of human elicited features successfully increase the model accuracy by 18.22%.

1. Introduction

Location analytics is also known as geospatial analytics, location intelligence or spatial intelligence. It is a process of transforming geospatial information into valuable knowledge and give insights to solve problem. Visualization of geographical data is easier and useful to gain fruitful knowledge. For example, when the geographical data is plotted on a map, some information and trend will be easily being discovered.

Most of the companies provide location intelligence in the business context to bring a better outcome for an organization (e.g. ESRI, Galigeo, Euclid Analytics, CISCO, etc.). Generally, business data that consists of geographical component is competent as it provides useful information and significant insights not only for marketing strategy but also brings benefits to the whole company.

To date, geospatial analytics has been gaining attention commercially and academically. It has been widely used in multiple domains includes retail, banking, insurance, healthcare, automotive, courier, public safety, airport, manufacturing and energy. For example, studies reported in [1], [2], [3], [4], [5], [6] employed site selection for the sake of energy conservation. Using site selection system, researchers evaluated the best site for biogas plants, wind farm or solar farms to utilize the renewable energy resources for the development of a sustainable environment. Likewise, site selection can be used in agriculture, Mishra and colleagues used site selection to find the suitable areas for organic farming [7]. Beyond that, Chaudhary and colleagues as well as Kumar and Bansal used location intelligence for safety planning [8, 9]. Disaster planning makes use of location analytics to strengthen

the prevention efforts. In this paper, the research will focus on retail site selection and to determine the optimal features that are applicable for retail site analysis.

In light of the standpoint from retail and business, location is a key factor for a business to run successfully in long-term. A prime location decision is imperative to achieve maximum financial gains as well as to reinforce the development of the business. A good location decision is able to increase the customer markets and has a great impact on market share. However, to determine the placement for a new installment of the business is both time consuming and labour intensive. Currently, there is lack of information about the set of optimal features for telecommunication company. In this paper, the objectives of the study are:

(i) to study whether one year sales data is sufficient to create a good predictive model
(ii) to study the impact of human elicited features on the predictive accuracy of the model

The rest of the paper is organized as follows: Section 2 reviews the related work about the features used in the retail site selection. Section 3 describes the datasets and method used in this research. Section 4 discusses the results of the experiment and Section 5 gives conclusion.

2. Related work

Recently there are extensive researches have been conducted on site selection. Studies shows that location is an important factor for site selection. Karamshuk and colleagues utilized the check-in data in Foursquare to identify the best retail store location from a set of potential candidates areas [10]. Researchers focus on the geographic features and user mobility to identify the popularity of an area ,using it as an index to determine the placement of the retail stores. The research shows that the most robust indicators of the popularity of an area are its location characteristics (e.g., train station or airport) and type of retail that is same to the target business. Lin and colleagues proposed a new location analytics framework that utilized the ubiquitous of social network data to predict the popularity of a particular location, which reported could bring success of a business [11]. The research's findings disclose that the neighbor (location characteristic) of a business is a crucial factor that will influence the popularity of a business. Roig-Tierno and colleagues study geodemand and geocompetition to determine the location for new establishment [12]. The research used Analytics Hierarchy Process (AHP) to identify the main criteria, namely establishment, location, demographics and competition factors that will affect the performance a supermarket. The paper concluded that the most significant features that influence the successfulness of a supermarket are those that are related to location and competition. In addition, a study by Erbıyık and colleagues took into account of the costs, competition, traffic density, physical features and location as factors to determine the location of a retail store, e.g. a production plant in Konya, Turkey [13]. The study shows that the most important criteria are traffic density and competition conditions.

On the other hand, Suárez-Vega and colleagues used competitive models to figure out the most promising locations to locate a single facility in a franchise distribution system. The variables that the research studied were sales surface area, estimated capture and market share to build the competitive location model [14]. Besides, Turk, Kitapci and Dortyol emphasized the importance of consumption maps to determine the optimum locations [15]. The paper is a case study that uses the features such as population, average income and monthly consumption that are related to education, health and food to decide the locations for supermarket. Chen and Tsai considered the features in the aspect of demographics, market conditions, store conditions, accessibility conditions and sales performance [16]. The study concludes that the location factors, i.e. store size, availability of parking area, store visibility and population growth rate of the vicinity, are the most significant factors that will affect the performance of a retail store.

In addition, Yıldız and Tüysüz [17] employed Hesistant Analytic Hierarchy Process (H-AHP) and Gray Relational Analysis (GRA) for food retailing. They mentioned that location is strategically important for organized food retailers because the location of retail store will affect the overall success of a retailer. Chacón-García [18] applied analytical Geographic Information System(GIS) and AHP to identify the location for the new opening of pharmacy (retail). The main criteria used in their study are

areas of maximum profitability and neighborhood area. Their study shows that the combination of GIS and multicriteria decision method is efficient to measure the spatial reality and its influence in retail businesses. Apart from retail site selection, location intelligence was also applied by Cabello for locating bank branch. The author identified the external and internal factors that were used to decide a branch location [19]. External factors include unemployment, density of population, percentage of foreign population of an area; while internal factors are the characteristics features of each business entity. The proposed method minimizes the distance from the candidate-site' features to the successful bank branches with the purpose of imitating the features of existing branch that has good performance.

As shown above, most of the research [10, 11, 12, 16] concluded the location factor is most significant in affecting the performance of a business.

This study focused on the location, property, demographics, economic and educational features, in addition with five human elicited features, e.g. building type, visibility, accessibility by public transport, availability of parking space and type of access to entrance to determine a set of optimal features that is significant and applicable for retail site selection.

3. Methodology

3.1. Dataset

The six datasets used in this research were Point of Interest (POI), Yellow Pages (YP), property, population, economic and educational data. The demographics data was gathered from Department of Statistics Malaysia (DOSM) while the rest of the datasets were provided by Telekom Malaysia (TM) .

Let D_{poi} denotes the POI dataset, it records 418324 points of interest in Malaysia which is catogorized into 1232 categories (e.g. western restaurants, Starbucks, book store, fashion stores etc.). D_{poi} represents geographical data that consists of geospatial information of a place. Examples of attributes in D_{poi} are place name, POI code, longitude, latitude, city and state. POI code is the code that categorized the point of interest based on different categories.

The Yellow Pages dataset is represented as D_{yp}, it is a list of businesses in Malaysia. The list consists information about a company and its location details. Examples of attributes in D_{yp} are business name, YP code, longitude, latitude, city and state. YP code is the code used to categorize the businesses into 3081 types of business (i.e. beauty saloons, domestics services, laundries etc.).

Let D_{ppt} denote the property data, which it documents the type of property (i.e. apartments, town house, storey shop etc.) of a given location. Examples of attributes in D_{ppt} are property type, longitude and latitude.

Economic data, D_{eco}, tabulated the job type, industrial field and employment status of the workforce in Malaysia. There are 10 types of job, 22 types of industrial field and 5 levels of employment status in this study. The educational data, D_{edu} consists of the attributes that represent the status of schooling, education level and certificated achieved of the people in a particular area. There are 4 types of schooling status, 9 educational levels and 11 types of certificate recorded in D_{edu}.

Let D_{pop} denotes the population data. The data is collected based on different races population in Malaysia at the locality authority area or district level. The races included in the research are Malay, *other bumiputera*, Chinese, Indian, non-Malaysian citizens and others.

All the datasets above are automatically extracted for a particular store of a business. The algorithm of data extraction will be explained in the following section. Apart from the automated extract data, five extra features are also included in the research. The features are identified by the expert, which will affect the sales performance and the decision making of determining the location of a new establishment. The five features are building type, visibility, accessibility by public transport, availability of parking space and type of access to entrance. Due to granularity of the dataset, these human elicited features require manual collection, which is done through Google Street View.

Sales data was provided by one of the telecommunication company in Malaysia. The sales data will be the dependent variable that allows this work to create a predictive model based on the sales performance. The sales data provided are monthly sales in year 2015 and 2016. It will then aggregate

into yearly sales. This research aims to study whether one year sales data is sufficient for developing a satisfied predictive model. All the data are collected for the 96 stores of the telecommunication company.

3.2. Data Extraction

The process of data extraction extracts all the information needed for the 96 outlets. Let F_{eco} and F_{edu} be the economic and educational features. These features can be directly extract from the dataset because the data are tabulated for a particular area, i.e. district. *inner_join* function and *merge* function are used to match the data for the 96 outlets by their districts. Next, let F_{pop} be the population features. As described in the previous section, there will be 6 columns for F_{pop}. The population information is collected in the way that: (i) the population in the smallest granularity of area (local authority area) is searched and matched for the outlets, (ii) the population in the higher granularity of area (district) is searched for the remaining unmatched outlets.

On the other hand, F_{loc} are the features that represent the location characteristics of a given location, it needs to be extracted from D_{poi} and D_{yp}. Algorithm 1 is proposed to extract the location characteristics given a location.

Algorithm 1. $get - locationFeature$

Input: $D_{outlet}, D_{poi}, D_{yp}$
Output: $Floc$
1: ***for*** $each\ point\ in\ D_{outlet}$ ***do***
2: $d_{poi}^{100} \leftarrow D_{poi} \leq 100m$
3: $d_{yp}^{100} \leftarrow D_{yp} \leq 100m$
4: $d_{poi+yp} \leftarrow \left(d_{poi}^{100} \cup d_{yp}^{100}\right) \leq 100m$
5: $d_{duplicate} \leftarrow match\left(d_{poi+yp}\right)$
6: $F_{loc} \leftarrow \left(d_{poi}^{100} \cup d_{yp}^{100}\right) - d_{duplicate}$
7: ***end for***

Algorithm 1 shows the data extraction process to extract the location features. Let D_{outlet} be the list of outlets or branches of a business that consists of location details such as longitude and latitude. First of all, the algorithm is constructed to obtain the nearby features (F_{loc}), e.g. shops, restaurants, hotels or businesses around the outlet. From both D_{poi} and D_{yp}, the nearby features, d_{poi}^{100} and d_{yp}^{100} are extracted within $100m$ radius surrounding an outlet. However, d_{poi}^{100} and d_{yp}^{100} cannot be combined directly to get the final F_{loc}. This is caused by the overlapping infomation in both datasets. Let $d_{duplicate}$ be the set of features that are duplicating in d_{poi}^{100} and d_{yp}^{100}. In order to get $d_{duplicate}$, the algorithm runs the unification of d_{poi}^{100} and d_{yp}^{100} with the radius of $10m$. This is to find two similar points in d_{poi}^{100} and d_{yp}^{100} that share the same information even though they have slightly different naming and coordinates. Next, $d_{duplicate}$ is removed from the unification of d_{poi}^{100} and d_{yp}^{100} to get rid of the replication and the final F_{loc} is obtained.

Let F_{ppt} denotes the property features. Similarly, F_{ppt} is extracted from D_{ppt} using the same technique as line 2 in the algorithm. Algorithm 1 (line 2) is aimed to search the nearby features (location or property) within $100m$ radius in the dataset.

Let F_{hum} represents the human elicited features. As mentioned in the section previously, the five human elicited features are collected manually through Google Street View for the 96 outlets for the telecommunication company.

3.3. Data Preprocessing

Data preprocessing is a crucial step to obtain a model with outstanding performance. In this research, data cleaning is carried out on D_{poi} and D_{yp} to ensure the matching function can run smoothly to get a set of reliable F_{loc}. Elimination of punctuations and capitalization are executed on the attributes of place name in D_{poi} and business name in D_{yp} to increase the capability of matching process.

Besides, discretization is performed on the sales data as the research intends to classify the sales into three categories, namely, low, moderate and high.

3.4. Analytics Dataset

After obtaining all the features, the features are then combined to form an analytics dataset, a data frame that is ready for the subsequent analysis tasks. In this research, three analytics were formed in order to study the objectives of the paper.

The first analytics dataset, AD_1 consists of 30 F_{loc}, 26 F_{ppt}, 34 F_{eco}, 22 F_{edu}, and 6 F_{pop}, which is an analytics dataset with total of 118 attributes. The sale in year 2016 is then added to AD_1. AD_1 will acts as the control experiment dataset, which is used to compare with the other two analytics datasets to study the objectives.

Let AD_2 indicates the second analytics dataset. The attributes of AD_2 are the same as AD_1. It consists of 118 features. In AD_2, the summation of sales in year 2015 and 2016 is aggregated into it. AD_2 is used to study the first objective, to see whether one year of sale data is sufficient for the model.

The third analytics dataset, AD_3 consists of the 118 features with addition of the five human elicited features, F_{hum}. Therefore, AD_3 will have 123 features. Sale data in 2016 is combined with AD_3. AD_3 is used to compare with AD_1 to investigate the impact of human elicited features on the predictive model.

3.5. Selecting the Optimal Feature Set

In this research, the analytics datasets obtained are high dimensional data, which the number of features (118 or 123) is largely more than the number of observations (96). Therefore, feature selection is employed in this research. Feature selection is used to alleviate the curse of dimensionality coined by Bellman [20]. Curse of dimensionality is a condition where the high-dimensional spaces do not met in low-dimensional settings. High-dimensional spaces will incur Hughes phenomenon, where when then dimensionality of the data increases, the predictive power decreases [21]. In this light, feature selection is implemented to obtain a subset from the set of initial features that will retain enough information to achieve sufficiently satisfied results. In this research, 6 feature selection methods are used, namely, Boruta, Hill Climbing Search (HCS), Random Search (RS), Thresholding (THRES), random forest - mean decrease impurity (RF-IM), and random forest - mean decrease accuracy (RF-AC). The types of the feature selection methods are shown in Table 1.

Table 1. Types of feature selection

Feature Selection	Type	Description
Boruta	Wrapper	Wrapper method is like a search problem. In this method, a subset of features is used to train a model. Different combination of features and model is evaluated and compared. Eventually, the best subset will be selected based on the model accuracy.
Hill climbing search (HCS)	Wrapper	
Random Search (RS)	Wrapper	
Thresholding (THRES)	Feature ranking	Feature ranking calculates the importance or performance of each feature and then ranks the features in descending order. User needs to determine the cutoff point or how much features to retain for the analysis.
Random forest - mean decrease impurity (RF-IM)	Feature ranking	
Random Forest - mean decrease accuracy (RF –AC)	Feature ranking	

Boruta is a wrapper algorithm around random forest that helps in understanding the mechanisms related to the predictors. It can be implemented using the R package named *Boruta*. Hill climbing is an algorithm that starts with a random feature set, and then proceeds to evaluates the features to choose a best subset of features. The relevant R package is *FSelector*. The next feature selection algorithm is random search. It can be applied through the *mlr* package. Random search allows the algorithm to move to avoid it stays on a local minimum. Thresholding step is a step of variable selection in VSURF package that eliminates the irrelevant features in the datasets. Random forest provides two methods for feature selection: (i) mean decrease impurity and (ii) mean decrease accuracy. Impurity in random forest is a measure when the optimal condition is chosen. The features are ranked based on how much they decrease the weighted impurity when training a tree. Mean decrease accuracy is a feature selection method that computes the impact of each feature on the model accuracy. It permutes the values of each variable and calculates how much the permutation decreases the model accuracy. For instance, the permutation of insignificant feature will not have much impact on the accuracy of the model. These two feature selection methods are implemented through the *FSelector* package.

3.6. Predictive Model Construction

Since the objective of this study is to investigate the performance of datasets (one year or two year sales data and with or without human elicited features), the classifiers used should be not significant and concern of this research.

However, the current research employed 11 classifiers to obtain the predictive accuracy. The 11 classifiers are come from six families, specifically, Bayesian approach (4), random forests (1), decision trees (3), support vector machines (1), bagging (1) and boosting (1).

There were four Bayesian approaches used in this study, namely, greedy thick thinning (GTT), tree augmented Naïve Bayes (TANB), augmented Naïve Bayes (ANB) and Naïve Bayes (NB). These classifiers were implemented through GeNIe. The rest classifiers were implemented in R.

The three decision trees used were C4.5, conditional inference tree (CTREE) and classification and regression tree (CART). The support vector machines used is the Support Vector Machines with Linear Kernel (SVML). Random forest (RF), C4.5, CTREE, CART and SVML were applied through the *caret* package. The relevant package for bagging (BAG) is *ipred*. The boosting method used in this research is adaboost (ABO) where the relevant package is *adabag*.

In this research, the conventional validation (e.g. partitioning the analytics dataset into 80% for training set and the remaining 20% for testing purpose) was used to evaluate the performance of each feature sets.

4. Results and Discussion

In this research, 6 feature subsets were obtained for each of the analytics datasets. Each of the feature subset was used to estimate the sales using the 11 classifiers. The results of the experiment are presented in the Table 2-4.

A t-test was conducted to investigate the different between AD_1 and AD_2. $p-value = 0.01976$ indicates that there is a significant different between the analytics datasets. The average accuracy of AD_2, 55.03% is significantly higher that the average accuracy of AD_1, 52.02%. The model accuracy increases by 3.01% when using two years sales data. This can be concluded that two years of sales data is able to produce higher accuracy than one-year sale data.

Besides, the other t-test was conducted to study the impact of the human elicited features on the model accuracy. $p-value < 2.2e-16$ indicates there is a significant difference between the analytics datasets with and without human elicited features. The average accuracy of AD_3, 70.24% is significantly higher that the average accuracy of AD_1. After adding the human elicited features, the model accuracy is successfully increases by 18.22%. This can be concluded that the human elicited features bring significant increases for the model accuracy.

Besides, the highest predictive accuracy is 83.3% obtained from the feature selection RF-IM and classifier C4.5. The features retained in RF-IM for AD_3 are malay food restaurants, shopping shop,

banks (F_{loc}), human health and social work activities, skilled worker and carpenters, self-employed, public administration and defense, electricity, gas, steam and air conditioning supply (F_{eco}), low secondary education, not in school, never go to school, PMR, low education, UPSR, still in school (F_{edu}), five storey house, low cost house (F_{ppt}), availability of public transport, building type and type of access to entrance (F_{hum}).

Table 2. Predictive accuracy of AD_1 (%)

Feature selection	Bayesian approach				RF	Decision trees			SVM	bagging	boosting
	GTT	TANB	ANB	NB	RF	C4.5	CTREE	CART	SVML	BAG	ABO
Boruta	45.9	49.5	50.5	50.5	52.4	57.1	42.9	42.9	57.1	52.4	52.4
HCS	58.6	55.9	50.5	46.8	47.6	42.9	42.9	47.6	42.9	52.4	47.6
RS	55.0	48.6	40.5	40.5	57.1	57.1	47.6	52.4	52.4	57.1	57.1
THRES	64.9	54.1	51.4	49.5	57.1	66.7	33.3	38.1	57.1	61.9	38.1
RF-IM	62.2	63.1	44.1	50.5	61.9	57.1	38.1	38.1	52.4	42.9	52.4
RF-AC	64.0	64.0	62.2	48.6	52.4	57.1	52.4	57.1	66.7	66.7	52.4

Table 3. Predictive accuracy of AD_2 (%)

Feature selection	Bayesian approach				RF	Decision trees			SVM	bagging	boosting
	GTT	TANB	ANB	NB	RF	C4.5	CTREE	CART	SVML	BAG	ABO
Boruta	51.4	51.4	60.4	60.4	61.9	61.9	61.9	61.9	61.9	61.9	61.9
HCS	47.7	62.2	46.8	45.9	52.4	52.4	52.4	61.9	38.1	52.4	61.9
RS	52.3	62.2	55.9	46.8	52.4	52.4	52.4	57.1	57.1	47.6	57.1
THRES	59.5	68.5	62.2	55.9	52.4	52.4	52.4	61.9	52.4	52.4	38.1
RF-IM	64.9	65.8	51.4	48.6	52.4	57.1	52.4	61.9	42.9	57.1	52.4
RF-AC	59.5	62.2	55.9	54.1	57.1	52.4	52.4	61.9	52.4	38.1	47.6

Table 4. Predictive accuracy of AD_3 (%)

Feature selection	Bayesian approach				RF	Decision trees			SVM	bagging	boosting
	GTT	TANB	ANB	NB	RF	C4.5	CTREE	CART	SVML	BAG	ABO
Boruta	69.5	64.2	65.3	65.3	77.8	77.8	77.8	77.8	77.8	72.2	77.8
HCS	71.6	72.6	62.1	50.5	66.7	55.6	72.2	72.2	72.2	72.2	61.1
RS	70.5	70.5	62.1	53.7	72.2	61.1	61.1	66.7	66.7	77.8	72.2
THRES	80.0	71.6	70.5	52.6	66.7	72.2	72.2	72.2	77.8	72.2	72.2
RF-IM	71.6	74.7	66.3	60.0	72.2	**83.3**	77.8	77.8	77.8	77.8	77.8
RF-AC	69.5	65.3	58.9	55.8	77.8	66.7	77.8	77.8	66.7	77.8	77.8

5. Conclusion

In short, this research used six feature selection methods and 11 classifiers to build the predictive model that can estimate the sales given a location. From the results shown, it can be concluded that one-year sale data is not sufficient to build a satisfactory predictive model. The addition of sales data and the human elicited features is proven useful in increasing the model accuracy. From the features retained by the analytics dataset of highest accuracy, it can be shown that population is not selected and does not give impact on the model accuracy. This may be caused by the granularity of population collected is not the smallest granularity, i.e. town level that can most represent the people in an area.

The model developed in this research can be used to save time and cost as well as the labour effort to investigate the location for new openings. The selected features are used for retail site selection in this research which enables the telecommunication company to be able survey the locations that fulfill the feature characteristics to make decision for the new establishments.

Acknowledgement

This research work presented in this paper is funded by Telekom Malaysia.

References

[1] Brewer J, Ames DP, Solan D, Lee R and Carlisle J 2015 Using GIS analytics and social preference data to evaluate utility-scale solar power site suitability *Renew. Energy* **81** 825-36
[2] Franco C, Bojesen M, Hougaard JL and Nielsen K 2015 A fuzzy approach to a multiple criteria and Geographical Information System for decision support on suitable locations for biogas plants *Appl. Energy* **140** 304-15
[3] Latinopoulos D and Kechagia K 2015 A GIS-based multi-criteria evaluation for wind farm site selection. A regional scale application in Greece *Renew. Energy* **78** 550-60
[4] Mari R, Bottai L, Busillo C, Calastrini F, Gozzini B and Gualtieri G 2011 A GIS-based interactive web decision support system for planning wind farms in Tuscany (Italy) *Renew. Energy* **36(2)** 754-63
[5] Shaheen M and Khan MZ 2016 A method of data mining for selection of site for wind turbines *Renew. Energy* **55** 1225-33
[6] Uyan M 2013 GIS-based solar farms site selection using analytic hierarchy process (AHP) in Kara- pinar region, Konya/Turkey *Renew. Sustain. Energy Rev.* **28** 11-7
[7] Mishra AK, Deep S and Choudhary A 2015 Identification of suitable sites for organic farming using AHP & GIS *Egypt. J. Remote Sens. Space Sci.* **18(2)** 181-93
[8] Chaudhary P, Chhetri SK, Joshi KM, Shrestha BM and Kayastha P 2016 Application of an Analytic Hierarchy Process (AHP) in the GIS interface for suitable fire site selection: A case study from Kathmandu Metropolitan City, Nepal *Socioecon. Plann. Sci*. **53** 60-71
[9] Kumar S and Bansal VK 2016 A GIS-based methodology for safe site selection of a building in a hilly region *Front. Archit. Res.* **5(1)** 39-51
[10] Karamshuk D, Noulas A, Scellato S, Nicosia V and Mascolo C 2014 Geo-spotting: mining online location-based services for optimal retail store placement *Proc. of the 19th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining* (ACM) pp 793-801
[11] Lin J, Oentaryo R, Lim EP, Vu C, Vu A and Kwee A 2016 Where is the Goldmine?: Finding Promising Business Locations through Facebook Data Analytics *Proc. of the 27th ACM Conf. on Hypertext and Social Media* (ACM) pp 93-102
[12] Roig-Tierno N, Baviera-Puig A, Buitrago-Vera J and Mas-Verdu F 2013 The retail site location decision process using GIS and the analytical hierarchy process *Appl. Geogr*. **40** 191-8
[13] Erbıyık H, Özcan S and Karaboğa K 2012 Retail store location selection problem with multiple analytical hierarchy process of decision making an application in Turkey. *Procedia-Soc. Behav. Sci.* **58** 1405-14
[14] Suárez-Vega R, Santos-Peñata DR and Dorta-González P 2012 Location models and GIS tools for retail site location *Appl. Geogr.* **35(1-2)** 12-22
[15] Turk T, Kitapci O and Dortyol IT 2014 The usage of Geographical Information Systems (GIS) in the marketing decision making process: a case study for determining supermarket locations. *Procedia-Soc. Behav. Sci.* **148** 227-25
[16] Chen LF and Tsai CT 2016 Data mining framework based on rough set theory to improve location selection decisions: A case study of a restaurant chain *Tour. Manag.* **53** 197-206
[17] Yıldız N and Tüysüz F 2018 A hybrid multi-criteria decision making approach for strategic retail location investment: Application to Turkish food retailing *Socioecon. Plann. Sci.*
[18] Chacón-García J 2017 Geomarketing techniques to locate retail companies in regulated markets. *Australasian Mark. J. (AMJ).* **25(3)** 185-93
[19] Cabello JG 2017 A decision model for bank branch site selection: Define branch and do not deviate *Socioecon. Plann. Sci.*
[20] Bellman R 2013 *Dynamic programming* (Courier Corporation)
[21] Hughes G 1968 On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inf. Theory* **14(1)** 55-63