Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2021) 29: 1509 – 1523

doi:10.3906/elk-2005-202

Determining and evaluating new store locations using remote sensing and machine learning

Berkan HÖKE[1,*], **Zeynep TURGAY**[1], **Cem ÜNSALAN**[2], **Hande KÜÇÜKAYDIN**[3]

[1]Migros R&D, İstanbul, Turkey

[2]Department of Electrical and Electronics Engineering, Faculty of Engineering, Marmara University, İstanbul, Turkey

[3]Department of Industrial Engineering, MEF University, İstanbul, Turkey

Received: 29.05.2020 • **Accepted/Published Online:** 19.01.2021 • **Final Version:** 31.05.2021

Abstract: Decision making for store locations is crucial for retail companies as the profit depends on the location. The key point for correct store location is profit approximation, which is highly dependent on population of the corresponding region, and hence, the volume of the residential area. Thus, estimating building volumes provides insight about the revenue if a new store is about to be opened there. Remote sensing through stereo/tri-stereo satellite images provides wide area coverage as well as adequate resolution for three dimensional reconstruction for volume estimation. We reconstruct 3D map of corresponding region with the help of semiglobal matching and mask R-CNN algorithms for this purpose. Using the existing store data, we construct models for estimating the revenue based on surrounding building volumes. In order to choose the right location, the suitable utility model, which calculates store revenues, should be rigorously determined. Moreover, model parameters should be assessed as correctly as possible. Instead of using randomly generated parameters, we employ remote sensing, computer vision, and machine learning techniques, which provide a novel way for evaluating new store locations.

Key words: Remote sensing, machine learning, competitive facility location, revenue estimation, utility model

1. Introduction

Development in internet technologies and change in urban mobility of metropols have greatly influenced the retail industry. Food retail industry is one of the areas, which has been deeply affected by these rapid changes. In particular, the demand for smaller stores rather than hypermarkets is growing due to the metropolitan lifestyle. This forces supermarket chains to open more boutique-type stores, with limited floor area, at the center of populated regions in addition to hypermarkets located at remote regions.

Boutique-type stores are generally designated as modern convenience stores to meet local customer demands by providing fast shopping opportunity. For this reason, they are preferred to be located in residential areas within walking distance. It is anticipated that number of such stores will increase due to latest tendency. Thus, choosing the right location for a new store is becoming more important. It is known that store location decisions are rare decisions that an organization must make. However, they are strategic and significant decisions since they create big investment cost, which cannot be easily compensated. Moreover, they determine the competitive advantage to the organization [1]. Hence, identifying the store location is of great importance for companies in food retail industry. Determining and evaluating potential store locations should be rigorously carried out by companies. On the other hand, the whole process of opening a new store should also be put into

*Correspondence: berkanh@migros.com.tr

practice as quickly as possible. The reason for this is that not every available location is substantially adequate to serve as a convenience store and competition for such available locations is high because of increasing demand.

In literature, studies on store location determination usually concentrate on mathematical models for locating large-scale facilities, where the demand is aggregated at macro-scale nodes which can either represent a city or town. Parameters of such models are the distance between the potential facility and demand nodes and the annual buying power at demand nodes. Moreover, parameters fed to mathematical models are of particular importance, since different parameters can lead to solutions which are very diverse. The distance parameter is usually expressed using conventional metrics, such as the Euclidean or rectilinear distance in literature. However, convenience stores are small-scale facilities targeted for neighborhoods rather than cities or towns. Therefore, it is crucial to estimate street walking distances, the total demand, and the demand distribution within the neighborhood as correctly as possible. Moreover, parameters such as walking distance of a store site to access public transport and the number and type of other facilities surrounding the store site are critical factors which specify the attractiveness of a store site.

The aim of this study is to present a specialized decision support system (DSS) for food retail industry in order to determine and evaluate new locations for small stores. The DSS chooses the most appropriate utility model among variety of existing models in literature, where the aim is to find the model that represents the conditions for food retail industry the best. In order to understand which model is the best one, several utility models are taken into account to compute the revenue of Migros Inc., one of the major supermarket chains in Turkey, using its historical data. Hence, the model inferring the revenue most accurately is selected so that the validation of the model is performed by the DSS. Moreover, the required parameters such as the location of demand points, the population at demand points, the distance between potential store locations, and the demand points are provided by the building detection algorithm from aerial and satellite images are developed within the DSS.

There exist several important distinctions of this paper from other studies especially in the CFL literature. First of all, the critical parameters, which significantly affect the revenue estimation such as the total demand and the demand distribution within a neighborhood are estimated as accurately as possible by employing remote sensing through stereo/tri-stereo satellite images rather than using randomly generated parameters, which is very common in the CFL literature. By this means, we manage to obtain logical values for these critical parameters. As a summary, the outputs of remote sensing process are used as input to the utility models which estimate the revenues that can be collected by the stores. In this way, an interdisciplinary research study is conducted to solve a complicated real-life problem. Secondly, a number of different utility models existing in the CFL literature are considered to explain the customer behavior generating store revenues using historical data of existing stores of Migros Inc. To the best of our knowledge, such a validation which generates managerial insights has never been done before for the food retail industry. By this means, the most suitable utility model, which estimates the revenue as closely as possible for a supermarket chain can be used in an optimization model to find out the best locations even for several new stores at the same time.

From the remote sensing or computer vision perspective, there are several studies on object (or building) detection from satellite and aerial images. However, this is the first study taking into account the actual problem (store location determination), while detecting buildings in crowded residential regions. Therefore, the developed methods for this purpose both benefit from the clues from the problem itself and become challenging based on the studied sites and available scenarios. In the following sections, we provide the proposed computer vision methods to satisfy both constraints.

The remainder of the paper is organized as follows: First, a thorough literature survey on the problem

is provided. Then, methods are introduced to estimate demand nodes via remote sensing data. Afterwards, revenue estimation based on the estimated demand nodes are investigated. Finally, experiments are provided followed by conclusions.

2. Literature review

The CFL problem is considered in this study, since it involves the competition between stores of the target company (Migros Inc.) and stores of its competitors in the market. The CFL problem can be further partitioned into two classes as nonreactive and reactive. In the nonreactive competition, it is assumed that competitors with existing facilities in the market do not react to opening of new facilities. In the reactive competition, the possible reaction of competitors is also taken into account [2]. Therefore, the reactive CFL problem tackles the issue of reaction by considering each competitor as a player who acts either sequentially or simultaneously. This can be formulated by a game-theoretic approach. However, in this study, the utility model validation is done using historical data, where the results of reactions of competitors are already observed, the location and type (design) of existing and new stores are already fixed, and the resulting revenue is known with certainty. Thus, a nonreactive CFL model serves the purpose of this study better than the reactive model.

The book by Farahani and Hekmatfar [3] is one of the recent studies, which scrutinizes the concepts, algorithms, and applications in facility location. Mendes and Themido [4] analyze the problems encountered in facility location at strategic and operational level and the criteria to evaluate the facility sites. Roig-Tierno *et al.* [5] determine the location for a new supermarket in Murcia, a city in Spain, by introducing a geographic information system utilizing analytical hierarchy process, which is able to identify critical factors in survival of supermarkets.

It is possible to divide CFL problems into two broad categories based on the utility function employed in the model as deterministic and probabilistic (random) utility models [2]. In both categories, the utility of a facility for a customer is calculated as a function of the facility attractiveness and distance between the customer and facility. The difference between these two types is that in deterministic utility models, customers visit only the facility, which provides the highest utility for them; whereas, in probabilistic utility models, customers visit each facility with certain probability. Three remarkable examples of deterministic utility models belong to Drezner [6], Plastria and Carrizosa[7], and Pelegrín *et al.*[8]. The first two studies try to find the optimal location for a single facility in continuous plane. However, the study by Plastria and Carrizosa [7] also determines the best attractiveness for the facility. The work by Pelegrín *et al.* [8] considers the expansion problem of a chain of facilities by opening a certain number of new facilities in the market with competing chains offering the same product or service, where customers choose a single facility providing the maximum utility for them. However, opening new facilities leads to the market share reduction for existing facilities of the expanding chain. This is the so-called cannibalization effect. In order to determine the optimal location for new facilities, a bi-objective optimization model is formulated.

The most frequently used probabilistic utility model in the CFL literature is the gravity-based model which is first introduced by Reilly [9] and later extended by Huff [10, 11]. In this probabilistic utility model, it is assumed that the probability that a customer visits a facility is proportional to the attractiveness of the facility and inversely proportional to the distance between the facility and customer. Later, Nakanishi and Cooper [12] proposed another probabilistic utility model referred as the multiplicative competitive interaction (MCI) model, where the utility is computed by taking the product of all facility attributes including the distance after each is raised to a power. Achabal *et al.* [13] determine the optimal location and design of a number

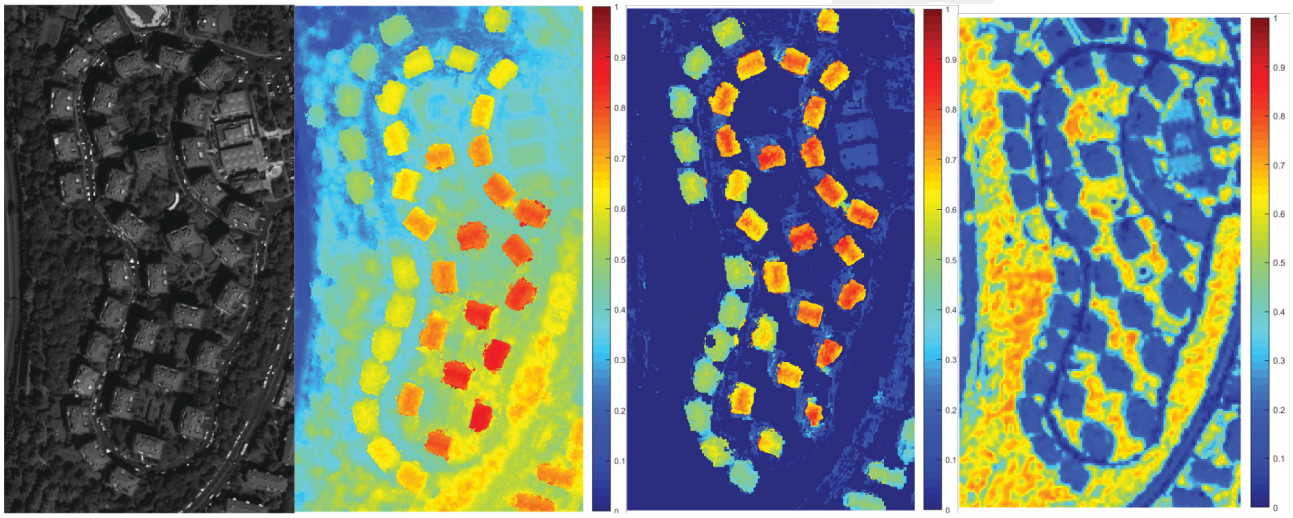
of new stores in discrete space which maximizes the market share by employing the MCI model. The model developed by Aboolian *et al.* [14] is similar to the model in [13]. The only difference is that Aboolian *et al.* [14] integrate the design properties into the model by making store attractiveness levels as continuous decision variables. However, they set a number of discrete alternatives for store designs while solving the model.

Random utility models are actually discrete choice models used for short-term travel decisions, where decision makers have a finite number of alternatives to choose. In CFL problems, which employ random utility models, the alternatives correspond to facilities/stores and decision makers correspond to customers choosing among facilities/stores. In such discrete choice models, utilities are in fact random variables which consist of a deterministic and a random part, where the random part comes into existence due to the unobserved characteristics of customers and stores. These random parts are commonly expressed by multinomial logit (ML) models [15]. Abouee-Mehrizi *et al.* [16] determine the optimal location, capacity of retail facilities and the price for providing the service from these facilities by representing the customer behavior utilizing the ML model. To this end, the probability of choosing a retail facility is calculated using an exponential function of demand limits and distance to facilities. All studies in the sequel employ the gravity-based model unless mentioned otherwise. Küçükaydın *et al.* [17] address a non-reactive CFL problem of a firm wishing to locate new facilities in a market with already existing facilities owned by competitors with the aim of maximizing the profit. For this reason, they find the optimal location, attractiveness levels, and number of new facilities in discrete space. Aboolian *et al.* [18] introduce an optimization model for non-reactive CFL problems which aim to locate certain number of facilities in discrete space by employing probabilistic utility models based on the gravity-based model. The developed model is a non-separable concave maximization knapsack problem which can be solved by a tangent line approximation procedure. Drezner and Drezner [19] examine a CFL problem in which the demand for a facility is affected by the customers' reluctance to visit the facility. They utilize the gravity-based rule by calculating the utility of a customer as an exponential decay function which increases with increasing attractiveness level. Benati and Hansen [20] make use of a utility function consisting of a deterministic and a random part. The deterministic part is a linear function of facility attractiveness and distance, whereas the random part follows Gumble distribution. The CFL problem in the continuous plane for locating one facility is addressed by Bello *et al.* [21]. They employ a minimax regret function to model uncertainty in the buying power of customers and solve the problem using a branch-and-bound method with a DCM bounding mechanism. Another noticeable study using the gravity-based rule belong to Drezner and Drezner [22]. Unlike the previously discussed studies, they do not locate new facilities. Instead, they try to infer attractiveness levels of existing retail facilities by making use of historic data including the realized sales figures and buying power of customers in communities. Drezner *et al.* [23] ignore the traditional assumption of fixed attractiveness levels. They rather assume that facility attractiveness levels are randomly distributed in the sense that customers have varying perception for the design of facilities and develop two solution procedures for locating p number of facilities. A detailed overview of CFL models can also be found in [24]. This study categorizes the CFL problems using seven components, namely decision variables, competition type, solution space, customer behavior, demand type, number of new facilities to locate, and the possibility for relocation and redesigning of existing facilities.

3. Estimating demand nodes using remote sensing

Satellite images are utilized for topography, urban, and vegetation analysis for decades. One of the tasks that is performed with satellite images is extraction of man-made objects, such as buildings. As a side note,

boutique-type stores are supposed to meet the demand of local customers. Hence, being close to demand nodes is crucial for retail companies in order to maximize revenue. In this study, we take buildings to be extracted from satellite images as demand nodes. To do so, we benefit from stereo images and multispectral information via computer vision methods. Stereo satellite images are the pair of images that are captured with different angles to involve area of interest. This process allows us to generate three dimensional terrain of the area by using 3D reconstruction techniques [25]. A sample satellite image from İstanbul is illustrated in Figure 1. Here, dark blue refers to the lowest and red color refers to the highest pixel values in the image, namely surface elevation.



(a) Satellite image (b) DSM (c) nDSM (d) NDVI

Figure 1. Information derived from satellite imagery. PLEIADES © CNES 2016, Distribution Airbus DS

Since each building is considered as a demand node, its area and number of storeys provide insight about the population, and hence the node size. To extract this information, we benefit from two well-known computer vision algorithms as mask R-CNN and semi-global block matching. Mask R-CNN is a convolutional neural network performing semantic segmentation which labels the building pixels on the image. The semi-global block matching (SGBM) algorithm seeks the matches between keypoints of a pair of images in order to reconstruct 3D representation of them. In this study, we use mask R-CNN and SGBM to extract building (demand node) information in line with prior information such as floor height and building area. Therefore, we can overcome the difficult problem of building detection and reconstruction in crowded residential regions. Furthermore, a map indicating node densities is generated by fusing the output of mask R-CNN and SGBM algorithms. Therefore, the obtained data in this stage is fed to the decision mechanism by revealing the magnitude of the demand node.

3.1. Estimating building areas

Convolutional neural networks (CNN) is a deep learning algorithm, which takes image as input and consists of learnable weights. Recently, CNN has become state of the art object detection method by outperforming conventional computer vision algorithms on several tasks such as the ImageNet large scale visual recognition challenge [26]. Capabilities of CNN have been extended by adding a region based object classification called

R-CNN [27]. Unfortunately, the elapsed time for training the whole dataset with R-CNN algorithm is quite long. Therefore, region proposal networks (RPN) and feature pyramid networks (FPN) are introduced within fast R-CNN and faster R-CNN algorithms to accelerate training times [28, 29]. Finally, a mask approach is fused with faster R-CNN to create the mask R-CNN structure which performs semantic segmentation on the input image [30]. Therefore, boundaries of an object can be represented with the help of this image mask.

In this paper, we benefit from the customized mask R-CNN implementation by Matterport [31] to detect building boundaries. This model is already trained with 280741 buildings most being from the United States. Unfortunately, properties of these buildings do not fit well with our problem region. Therefore, we had to apply transfer learning to recalculate the pretrained weights using 300 buildings from Istanbul, Besiktas area with 50 epochs. Hence, the weights of only the RPN, classifier and mask heads of the network is recalculated which has a significant time advantage compared to training all layers from scratch.

3.2. Estimating Building Heights

Besides building area, building height information must also be known to estimate the total volume of the demand node. LiDAR or stereo pair of optical images must be used in order to fulfill this task. Stereo satellite imagery is advantageous in our case compared with LiDAR due to its availability across the country and its wide coverage area. 3D reconstruction from a pair of images is an intense task and it may have horizontal artifacts when not properly used [32]. SGBM proposed by Hirschmuller [33] overcomes these horizontal artifacts. In this study, we benefit from an open-source satellite image processing framework MicMac to perform SGBM [34]. Stages of 3D reconstruction are as follows. First, tie points on two or more images are extracted using scale invariant feature transform (SIFT) [35]. Approximate nearest neighbors (ANN) are employed to seek tie point matches among different images [36]. Displacement of a tie point indicates the distance to the camera sensor. Therefore, displacement of building pixels are greater with compared to the ground pixels. Finally, these displacements are calculated including its direction using SGBM to obtain a robust depth map. This depth map is called digital surface model (DSM) which includes ground and vegetation heights as well as building heights. Digital elevation model (DEM) which contains only the ground elevation of the map is generated by erosion and thresholding operations on the DSM image. Difference between these two results with the height of buildings and vegetation is found as $nDSM = DSM - DEM$.

Moreover, near-infrared band of the satellite imagery is used to label vegetation. Therefore, the normalized difference vegetation index (NDVI) is calculated in order to locate vegetation pixels as

$$NDVI = \frac{NIR - R}{NIR + R} \tag{1}$$

where NIR denotes near-infrared band and R denotes the red band. DSM, nDSM, and NDVI are demonstrated for a test region in Figure 1.

A threshold is applied to NDVI image to distinguish vegetation and concrete structures. Afterwards, NDVI is used as a mask to eliminate vegetation using

$$T = \begin{cases} 0, & \text{if } NDVI > \tau \\ 1, & \text{Otherwise} \end{cases} \tag{2}$$

Hence, a new depth map is formed by $H_b = T \times nDSM$. Here, H_b represents the building height image

where pixel values denote the height value. This will help us to determine the correct population and node size.

4. Revenue estimation

In order to understand which utility model is the most suitable one for food retail industry, we consider several existing utility models presented in CFL literature. We assess the attractiveness levels, which are functions of facility attributes, of existing stores of the target company (Migros Inc). In fact, assessing attractiveness levels equals to assessing the parameters used in the function of facility attributes. As mentioned in the previous sections, utility models are used to model and explain the customer behavior by approximating the probability that a customer visits a facility. Then, the revenue that a facility collects from one customer is obtained by taking the product of the probability of the customer for choosing a facility and the annual buying power of that customer. The total annual revenue is calculated by summing up the revenues of all customers. Hence, the utility models also approximate the revenue collected by a facility at the same time. In light of this, we benefit from the annual revenues of existing stores of the target company and the customers are assumed to be aggregated at the demand nodes. Since the attributes and revenues of existing stores are already known, parameters of each utility model can be estimated by a regression analysis if additionally the buying power at demand nodes and the distance between demand nodes and existing stores can be obtained. The buying power of customers is provided by the Turkish Statistical Institute (TÜİK) for only cities and districts in Turkey, but not for neighborhoods or smaller regions. Since the main focus of this study is on boutique-type stores which are similar to convenience stores, the demand nodes should correspond to smaller areas rather than to districts or cities to make a more accurate estimation. In this regard, we take each building in a neighborhood as a demand node and the estimated surface area and the number of storeys for the buildings are used as representatives of size (population) of demand nodes. Hence, the findings in the previous section provide the size of the demand nodes which we use as a proxy for the buying power. Moreover, the walking distance between the demand nodes and existing stores and the competitors' facilities in the vicinity of existing stores are obtained from Google Maps.

We take the gravity based model proposed by Küçükaydın *et al.* [17] as the base utility model. In this model, the utility of a store located at site i for a customer at demand node j is computed as $\frac{Q_i}{d_{ij}^2}$, where Q_i shows the attractiveness level of store at site i and d_{ij} the distance between site i and demand node j. Thus, the total utility of m stores of one's chain for demand node j is found by $\sum_{i=1}^{m} \frac{Q_i}{d_{ij}}$. If there are r competing stores each having the attractiveness level q_k, the probability P_{ij} that a customer at demand node j visits the stores with the attractiveness level Q_i at site i is calculated as

$$P_{ij} = \frac{\frac{Q_i}{d_{ij}^2}}{\sum_{i=1}^{m} \frac{Q_i}{d_{ij}^2} + \sum_{k=1}^{r} \frac{q_k}{d_{kj}^2}} \tag{3}$$

where d_{kj} represents the distance between the competing store at site k and demand node j. Finally, the revenue that the store located at site i collects from n demand nodes each with buying power h_j is computed as $\sum_{j=1}^{n} h_j P_{ij}$.

Calculation of these probabilities can be easily explained by an example. Suppose that the store of one's

own chain (store 1) has attractiveness Q_1 and there are four competing stores with attractiveness levels q_2, q_3, q_4, and q_5 and four demand nodes. Data used in this example is presented by Table 1. Each row of the table corresponds to a demand node. The first column shows the distance between store 1 and each demand node, whereas the second column displays the buying power of every demand node. The third column shows the distance between each demand node and the competing store(s) in the vicinity of store 1 and the corresponding demand node. The distance between competing stores with attractiveness levels q_2 and q_3 and demand node 1 are 50 meters and 400 meters, respectively. The competing store with attractiveness levels q_4 is 200 meters away from the second demand node, which is in turn 200 meters away from store 1. Finally, the third demand node gets service from both store 1 and the competing store with attractiveness q_5 and the distances are 500 meters and 100 meters, respectively.

Table 1. Example data.

Distance (m)	Buying power	Competing stores & their distance (m)
100	1000	50, 400
200	5000	200
500	2000	100

In this case, the probability that demand nodes 1, 2, and 3 visit store 1 is calculated as

$$P_{11} = \frac{4 \times 10^8 Q_1}{4 \times 10^4 Q_1 + 64 \times 10^4 q_2 + 156.25 q_3} \tag{4}$$

$$P_{12} = \frac{Q_1}{Q_1 + q_4} \tag{5}$$

$$P_{13} = \frac{Q_1}{Q_1 + 25 q_5} \tag{6}$$

Other utility models which we make use of for revenue estimation are listed in Appendix. For each used utility model, the calculation of visiting probabilities are exhibited.

As mentioned before, the facility/store attractiveness is actually a function of store attributes. This function can be linear or non-linear. The store attributes can also include the number of competing stores in the vicinity of the store in addition to design features such as floor size of the store, the availability of a bakery and/or butcher, the possibility for orders by phone, the availability of alcoholic beverages. If a store attribute is represented by A_p and its weight by x_p, the store attractiveness Q can be calculated using a linear function $\sum_p x_p A_p$. As an alternative to a linear function, a nonlinear function can also be used following the MCI model proposed in [12]. With regard to this model, the store attractiveness can be found by an exponential function of attribute weights $\prod_p A_p^{x_p}$. In order to approximate the revenue as correctly as possible, one has to estimate the weight x_p of each attribute A_p in the best possible way. Therefore, these weights can be considered as sensitivity parameters which need to be estimated as accurately as possible and this can be done by building linear and nonlinear regression models.

4.1. Estimation using linear regression

In order to make use of linear regression, we first omit the utilities of competitors' stores for customers, which are calculated using the attractiveness levels of competing stores and the distance between demand nodes and competing stores. Instead, we incorporate the number of competing stores which are located within the circles centered at stores of one's own chain with radii of 50 and 100 meters as a store attribute. Since the utility models are employed to model the customer behavior/choice, the demographic features of customers can have a great influence on store revenue. With this in mind, we use the district-based data including 26 demographic features provided by TÜİK. First, we carry out a bivariate analysis by examining the simple correlation for every parameter pair to find the representative parameters. We further perform the bivariate analysis and obtain final eight representative parameters which can influence the revenue. These eight parameters are then turned into categorical variables and their effects on average household consumption for food and non-alcoholic beverages are analyzed by using generalized linear model (GLM).

In order to improve the results obtained by linear regression, an intuitive clustering is implemented based on observations where each cluster has strong building and revenue correlation. Then, linear regression is applied for each cluster separately. Using leave-one-out cross validation method, linear regression provides 14% MAPE with the highest error ratio of 38%. The accuracy of our predictive model based on linear regression has also been improved by implementing the leave-one-out cross validation.

4.2. Estimation using nonlinear regression

When the utilities of competitors' stores for customers are taken into account as demonstrated in the base utility model and other utility models given in Appendix, a non-linear regression analysis is required to approximate the sensitivity parameters. Since the goal of regression is to minimize the sum of squared errors, the function to be optimized for each utility model is non-convex and non-unimodal even with one competing store. Therefore, numerical and/or heuristic methods should be applied to approximate the minimum of this non-linear optimization problem. To this end, we develop three solution methods, namely the gradient descent method, Nelder–Mead method, and genetic algorithm.

The gradient descent (GD) method is an iterative first-order algorithm which is designed for finding local minimum of an unconstrained optimization problem with a multi-variable, differentiable objective function. In every iteration, a step size has to be determined which gives the highest improvement in the objective function. To find the optimal step size, we apply the golden section search (GSS) and dichotomous search (DS) methods. They give rise two different versions for the gradient descent method. A third basic version of the algorithm is obtained by taking the same step size at every iteration.

The Nelder–Mead (NM) method, is another numerical method for finding the local minimum of a non-linear, multi-variable objective function. Unlike the gradient descent method, it makes use of function comparisons rather than derivatives and starts with $n+1$ points (vertices) of a simplex in n dimensions. The method generates a sequence of simplexes, for which the function values at vertices become smaller [37].

Genetic algorithms (GA) are metaheuristic algorithms which start from an initial population and iteratively generate new populations which replace the current one [38]. At each iteration, GA alternate diversification and intensification phases, where the diversification is enabled via implementing crossover operators and intensification by local search. In order to produce good children with crossover operators, first the binary tournament selection method is implemented which does not perform well. Then, two methods are used to improve the performance. In the first method, the parents are randomly selected among the best 1000 solutions. In the

second method, parents are randomly selected among the whole population. In this way, we ensure that both good and bad parents take part in the population which increases the diversification amount.

5. Experiments

In order to evaluate the performance of the proposed method, we perform several experiments in this section. To do so, we first handle the remote sensing part and analyze its performance as a stand alone operation. Then, we focus on revenue estimation operations. Finally, we analyze the effect of remote sensing operations on revenue estimation.

5.1. Performance analysis of remote sensing operations

In order to evaluate the remote sensing part of the proposed method, building ground truth dataset is constructed from Istanbul and Ankara, Turkey. This dataset consists of 11,317 buildings of which 7460 and 3857 are from İstanbul and Ankara, respectively. To note here, building formations in these two cities are different. Hence, the proposed method does not have any bias towards a specific building type. Moreover, the constructed dataset consists of area, height, and number of storeys information for each building.

We provide the obtained results with the introduced methods in the previous section as in Table 2. We also tabulate these results as percentages in Table 3. Here, ground truth data covering geometry and height of 7760 buildings for Istanbul and 3857 buildings from Ankara are purchased from GISLAB. Among these, 300 of the Istanbul buildings are used for transfer learning. As can be seen in these tables, the proposed method provide acceptable results in the estimation operation. Moreover, total building area in Ankara is 1.83 km^2. The proposed method estimates it as 1.85 km^2. Average building height for the dataset is 18.72 m. The method estimates it as 16.98 m. Total building cover area for Istanbul is 1.62 km^2. The proposed method estimates it as 1.55 km^2. Finally, average building height in the dataset is 14.89 m. It is estimated as 13.14 m by the proposed method. The results obtained in this section indicate that this information can be used in revenue estimation for target store location to be evaluated in the next section.

Table 2. Comparison of ground truth and estimated buildings.

Location	**# of buildings**	**# of estimated buildings**	**FP**	**FN**
Istanbul	7460	7204	454	333
Ankara	3857	4099	246	88
Total	**11317**	**11303**	**700**	**421**

Table 3. Comparison of ground truth and estimated buildings, in percentages.

Location	**Area Error (%)**	**Height Error(%)**	**FP**	**FN**
Istanbul	16.6	11.6	6.3	4.6
Ankara	16.0	14.8	6.0	1.5
Total	**16.4**	**12.7**	**6.2**	**3.5**

In experiments, both SGM and mask R-CNN algorithms are implemented using CUDA on an Intel Xeon E3-1535M processor, NVIDIA Quadro P4000 GPU, and 32 GB RAM. Execution time of both algorithms

depends on the image size, namely $\mathcal{O}(W \times H \times D)$ for SGM and $\mathcal{O}(W \times H)$ for mask R-CNN where W, H, D denotes width, height, and disparity range, respectively. For instance, a stereo pair with size of 20653×13304 pixels (each pixel corresponds to $0.5 \times 0.5m^2$) from Ankara region took 36 and 4 min for SGM and mask R-CNN, respectively. These results are acceptable for our method since revenue estimation is not a real-time operation. The user can tolerate such delays instead of manually labelling and estimating building information which is a tedious task. Moreover, the manual operation is prone to errors. However, the proposed method overcomes most of these problems by using automated computer vision methods on satellite images.

5.2. Performance analysis of revenue estimation operations

Output of automated vision algorithms, namely building locations, height, and area are employed as demand nodes, and the following revenue estimation methods utilize these as a store feature around the corresponding location. In other words, algorithms produce an approximate revenue using building specifications around 250 meters as input. In order to test revenue estimation, five algorithms are considered, namely the basic GD algorithm using the same step size, GD algorithm using GSS, GD algorithm using DS, NM method, and GA on the base utility model. Through experiments, population size is chosen 1000 for the GA algorithm which uses binary tournament selection. For the NM algorithm, the reflection coefficient is set equal to 1, the expansion coefficient to 2, the contraction coefficient to 0.5, and the shrink coefficient to 0.125. Experiments are run on a computer with Intel Core i7 - 7600U 2.80 GHz processor with 16 GB RAM. All algorithms are implemented in C#. Execution time for the algorithms is about 5 min. As a note, computation time is not the main concern here since implementation is not supposed to run on real-time.

The results are demonstrated using 20 nodes in Table 4. Note that there are three GD algorithms which are tested. Since the GD algorithm using GSS outperforms the GD algorithm using DS and the GD algorithm using the same step size, only the results obtained by the GD algorithm using GSS are given in Table 4. As can be seen in this table, the NM method outperforms the GA and the GD algorithm using GSS. When the average and maximum σ/μ values are considered, it can be realized that GD algorithms gets trapped in local minima, whereas the NM and GA manage to escape it. Therefore NM is chosen to be utilized for all the remaining utility models. In order to obtain sound results and completeness, we apply the same analysis to all of our stores covered by satellite images, which is totally 165 stores.

Table 4. Results with 20 demand nodes.

σ/μ	**NM**	**GA**	**GD (using GSS)**
Average	0.00654	0.01493	0.21249
Minimum	0.00002	0.00330	0.05582
Maximum	0.01493	0.06018	0.75570

The non-linear regression models are designed using the customer choice models already recommended in the literature as presented in Appendix. The models are developed with the Nelder–Mead algorithm and tested with synthetic data generated for small scale problems involving up to 20 nodes. The models are then tested with real life data and the mean absolute percentage error (MAPE) values are calculated as 1500%. High MAPE values indicate that these models are vulnerable to outliers and possible errors in data collection.

Next, the estimated building density is calculated within a radius and introduced to the linear regression models. The estimated building density is either calculated by the method described in Eqn. 3 or by a function

of estimated building and its distance to the candidate location. By means of the linear regression applied in separate clusters, where each cluster has a strong building and revenue correlation, our revenue estimation reaches MAPE values of 14%, when the multinomial logit model is selected as the utility model (see Eqn. 8).

5.3. The Effect of Remote Sensing Operations on Revenue Estimation

As the final set of experiments, we focus on the effect of remote sensing operations on revenue estimation. To do so, we run experiments on the exact same stores without using building information that we extract in Section 3. Then, we observe the effect of using demand node size obtained by remote sensing. The parameters that we use for estimating the revenue are the number of nearby stores (restaurants, cafe, clothing stores) retrieved from Google Maps, 8 demographic features obtained from Turkey Statistical Institute (TÜRKSTAT), and nearby buildings with height and area information. Nearby store features are the total number of surrounding facilities around 250 m of corresponding store location. Since we observe the improvements based on remote sensing contribution of this study in Section 3 and the employed features from TURKSTAT are confidential, the effect of remote sensing operations is analyzed. Hence, the revenue estimation is done both with and without remote sensing output. Through these analyses, we observe that 'nearby buildings' feature is very significant in terms of both regression and clustering steps in the proposed method. It decreases the mean absolute percentage error (MAPE) value from 26% to 14%. Hence, we can conclude the remote sensing step has a significant effect on revenue estimation and increases its accuracy.

6. Conclusion

It is a crucial task for retail companies to seek optimal store location, especially for boutique type stores, as they mostly need to be in range of customer walking distance. Therefore, the area with the highest population must be chosen on a new store opening decision. To do so, we propose a two step method. The first step of the proposed method is based on automated computer vision algorithms used in remote sensing applications. More specifically, we benefit from stereo satellite images with multispectral information with mask R-CNN and SGBM methods to detect buildings (with their area and height). In order to test the performance of the remote sensing step of the proposed method, we conduct several experiments. These experiments indicate that this part of the proposed method performs fairly well in diverse regions and building types. As the second step of the proposed method, we focus on revenue estimation using machine learning algorithms. The information extracted in the remote sensing step is used in this operation as well. Therefore, the proposed method benefits from the joint usage of remote sensing algorithms and revenue estimation methods. To be more precise, building location and height data is used as demand node size in the second step of the proposed method which consists of revenue estimation. This is the main novelty of the proposed method. We also conduct several tests to measure the performance of revenue estimation methods. Throughout these, we observe that the revenue estimation is accurate for most of the stores in our dataset including 165 stores. However, some stores still remain as outliers which prevents us from estimating the revenue of new stores within the expected accuracy. We finally perform experiments on measuring the effect of the remote sensing step on revenue estimation. These experiments clearly indicate that the remote sensing step positively affects the revenue estimation results. Hence, fusing techniques from remote sensing and industrial engineering disciplines improves the revenue estimation compared to using industrial engineering techniques alone.

We plan to extend this study with future work. First, we plan to improve the remote sensing part of the proposed method such that it affects the revenue estimation part better. Second, we can focus on

outlier/anomaly detection for stores with low expected revenue. Hence, they can be discarded from overall calculations. Third, we also plan to improve the linear regression accuracy and decrease error rates by replacing intuitive clustering with an algorithm such as k-means, EM or hierarchical clustering. Alternatively, sample consensus models such as RANSAC or MLESAC can be used to solve the clustering parameters for more accurate regression. The latter extension would be appending new attractiveness parameters to the model in order to enhance the estimation results. We believe that the present form of the study and its possible extensions emphasize the multidisciplinary characteristics of the proposed method here. Being such a multidisciplinary study, we expect it to be influential in both disciplines in future studies as well.

Acknowledgments

This research is funded by TÜBİTAK with project number 3151002. We would like to thank Yunus Emre Seyyar from TÜBİTAK and Prof. Selim Aksoy from Bilkent University for their valuable feedback during the project, Prof. Aytül Erçil from Vispera and Prof. Muhittin Gökmen from Divit for their valuable technical support and comments that improved our results, and finally Aydın Parmaksız and Kerim Tatlıcı from Migros Inc. for their managerial support of the project.

Some of the methods in this publication are patented under TURKPATENT with patent application number 2016/18104 dated on 08-12-2016 and TURKPATENT with patent application number 2020/07006 dated on 05/05/2020.

References

[1] Daskin M. Network and discrete location: models, algorithms and applications. Journal of the Operational Research Society 1997; 48 (7): 763–764.

[2] Küçükaydın H. Optimally Locating Facilities with Variable Characteristics. PhD, Boğaziçi University, Istanbul, Turkey, 2011.

[3] Farahani RZ, Hekmatfar M. Facility Location: Concepts, Models, Algorithms and Case Studies. Heidelberg,Germany: Springer, 2nd edition, 2009.

[4] Mendes AB, Themido IH. Multi-outlet retail site location assessment. International Transactions in Operational Research 2004; 11 (1): 1–18.

[5] Roig-Tierno N, Baviera-Puig A, Buitrago-Vera J, Mas-Verdu F. The retail site location decision process using gis and the analytical hierarchy process. Applied Geography 2013; 40: 191–198.

[6] Drezner T. Locating a single new facility among existing, unequally attractive facilities. Journal of Regional Science 1994; 34 (2): 237–252.

[7] Plastria F, Carrizosa E. Optimal location and design of a competitive facility. Mathematical Programming 2004; 2: 247–265.

[8] Pelegrín B, Fernández P, García Pérez MD. Profit maximization and reduction of the cannibalization effect in chain expansion. Annals of Operations Research 2016; 246: 57–75.

[9] Reilly WJ. The Law of Retail Gravitation. New York, USA: Knickerbocker Press, 1931.

[10] Huff DL. Defining and estimating a trading area. Journal of Marketing 1964; 28 (3): 34–38.

[11] Huff DL. A programmed solution for approximating an optimum retail location. Land Economics 1966; 42 (3): 293–303.

[12] Nakanishi M, Cooper LG. Parameter estimation for multiplicative competitive interaction model: least squares approach. Journal of Marketing Research 1974; 11 (3): 303–311.

[13] Achabal DD, Gorr WL, Mahajan V. Multiloc: A multiple store location decision model. Journal of Retailing 1982; 58 (2): 5–25.

[14] Aboolian R, Berman O, Krass D. Competitive facility location and design problem. European Journal of Operations Research 2007; 182 (1): 40–62.

[15] Hall RW. Handbook of Transportation Science. Kluwer Academic Publishers, 1999.

[16] Abouee-Mehrizi H, Babri S, Berman O, Shavandi H. Optimizing capacity, pricing and location decisions on a congested network with bulking. Mathematical Methods of Operations Research 2011; 74: 235–255.

[17] Küçükaydın H, Aras N, Altınel IK. A discrete competitive facility location model with variable attractiveness. Journal of the Operational Research Society 2011; 62 (9): 1726–1741.

[18] Aboolian R, Berman O, Krass D. Efficient solution approaches for a discrete multi-facility competitive interaction model. Annals of Operations Research 2009; 167: 297–306.

[19] Drezner T, Drezner Z. Lost demand in competitive environment. Journal of the Operational Research Society 2008; 59 (3): 362–371.

[20] Benati S, Hansen P. The maximum capture problem with random utilities: Problem formulation and algorithms. European Journal of Operational Research 2002; 143 (3): 518–530.

[21] Bello L, Blanquero R, Carrizosa E. On minimax-regret huff location models. Computers and Operations Research 2011; 38 (1): 90–97.

[22] Drezner T, Drezner Z. Validating the gravity-based competitive location model using inferred attractiveness. Annals of Operations Research 2002; 111: 227–237.

[23] Drezner T, Drezner Z, Zerom D. Competitive facility location with random attractiveness. Operations Research Letters 2018; 46 (3): 312–317.

[24] Ashtiani MG. Competitive location: a state-of-art review. International Journal of Industrial Engineering Computations 2016; 7 (1): 1–18.

[25] Hartley R, Zisserman A. Multiple View Geometry in Computer Vision. Cambridge, UK: Cambridge University Press, 2nd edition, 2004.

[26] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 2012; 25: 1097–1105.

[27] Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence 2015; 38 (1): 142–158.

[28] Girshick R. Fast r-cnn. In IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015; pp. 1440–1448.

[29] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 2017; 39 (6): 1137–1149.

[30] He K, Gkioxari G, Dollar P, Girshick R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017; pp. 2961–2969.

[31] Abdulla W. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017.

[32] Birchfield S, Tomasi C. Depth discontinuities by pixel-to-pixel stereo. International Journal of Computer Vision 1999; 35 (3): 269–293.

[33] Hirschmuller H. Accurate and efficient stereo processing by semi-global matching and mutual information. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, USA, 2005; pp. 807–814.

[34] Rupnik E, Daakir M, Deseilligny MP. Micmac–a free, open-source solution for photogrammetry. Open Geospatial Data, Software and Standards 2017; 2 (1): 14.

[35] Lowe DG. Distinctive image features from scale-invariant keypoints. International journal of computer vision 2004; 60 (2): 91–110.

[36] Indyk P, Motwani R. Approximate nearest neighbors: Towards removing the curse of dimensionality. In Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, STOC '98. Association for Computing Machinery, New York, NY, USA, 1998; p. 604–613.

[37] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes: The Art of Scientific Computing. New York, USA: Cambridge University Press, 1986.

[38] Talbi EG. Metaheuristics: From Design to Implementation. Hoboken, NJ, USA: John Wiley & Sons, 2009.

[39] Drezner Z. Facility location: a survey of applications and methods, volume 1. New York, USA: Springer-Verlag, 1995.

[40] Drezner T. Derived attractiveness of shopping malls. IMA Journal of Management Mathematics 2006; 17 (4): 349–358.

Appendix

Utility models

Multinomial logit model (without considering attractiveness)

$$P_{ij} = \frac{e^{-d_{ij}}}{\sum_{i=1}^{m} e^{-d_{ij}} + \sum_{k=1}^{r} e^{-d_{kj}}} \tag{7}$$

is stated out in [22, 39].

6.1. Multinomial logit model (considering attractiveness and distance parameter)

$$P_{ij} = \frac{\frac{Q_i}{e^{\beta d_{ij}}}}{\sum_{i=1}^{m} \frac{Q_i}{e^{\beta d_{ij}}} + \sum_{k=1}^{r} \frac{q_k}{e^{\beta d_{kj}}}} \tag{8}$$

where β is the distance decay parameter [19, 40].

6.2. Multinomial logit model (considering attractiveness and a different distance function)

$$P_{ij} = \frac{\frac{Q_i}{1+e^{\alpha+\beta d_{ij}+\gamma d_{ij}^2}}}{\sum_{i=1}^{m} \frac{Q_i}{1+e^{\alpha+\beta d_{ij}+\gamma d_{ij}^2}} + \sum_{k=1}^{r} \frac{q_k}{1+e^{\alpha+\beta d_{kj}+\gamma d_{kj}^2}}} \tag{9}$$

where α, β and γ denotes a positive constant, the distance decay parameteri and the power decay parameter, respectively. [22].

6.3. Multiplicative competitive interaction model

$$P_{ij} = \frac{\prod_{l=1}^{p} Q_{il}^{\beta_l} \times d_{ij}^{\beta}}{\sum_{i=1}^{m} (\prod_{l=1}^{p} Q_{il}^{\beta_l} \times d_{ij}^{\beta}) + \sum_{k=1}^{r} (\prod_{l=1}^{p} q_{kl}^{\beta_l} \times d_{kj}^{\beta})} \tag{10}$$

where β and β_l denote parameter for sensitivity of P_{ij} with respect to the distance and attribute l, respectively [12].