Price and spatial distribution of office rental in Madrid: a decision tree analysis *

Máximo Camacho[†1], Salvador Ramallo[2], and Manuel Ruiz Marín[3]

[1]University of Murcia

[2]University of Murcia

[3]Technical University of Cartagena

Abstract

In this paper, we assess the drivers of office rental prices in the municipality of Madrid with a sample of 4,721 offices in March, 2020. The estimation was performed using the decision tree approach, which was built with a random forest algorithm. This technique allows us to capture the strong nonlinear component in the relation between price and its drivers, mainly geospatial location. Through a stratified analysis, we find that the willingness to pay high rent in the center of Madrid is a feature of particular relevance to medium-sized offices. For different reasons, we also find some office clusters located far from the city center with high rent for both large and small offices.

Key words: spatial economics, random forest, nonlinear, offices.

*We thank Fernando Lopez for helpful comments and suggestions, and the support of grants PID2019-107192 GB-I00 (AEI/10.13039/501100011033). This study is part of the collaborative activities carried out under the program Groups of Excellence of the Region of Murcia, the Fundación Seneca, Science and Technology Agency of the Region of Murcia Project 19884/GERM/15. Data and codes that replicate our results are available from the authors' websites. All remaining errors are our responsability.

[†]Contact person: Máximo Camacho, University of Murcia, Faculty of Business and Economics, Department of Quantitative Analysis, 30100, Murcia, Spain. E-mail: mcamacho@um.es

1 Introduction

Positive externalities generated by the proximity to means of production are the basis of the term economic agglomeration, mentioned in Smith's (1776) and Marshall's (1890) classic texts. Duranton and Puga (2004) review the three microeconomic foundations present in the related literature to explain the concentration processes: *sharing*, referring to the possibility of sharing infrastructures and obtaining greater access to suppliers; *matching*, referring to a more efficient worker-company match; and *learning*, referring to the generation and spreading of new techniques. Alternatively, a Darwinian perspective based on company selection processes as an explanation for company concentration is proposed by Melitz (2003) and Melitz and Ottaviano (2008). Combes et al. (2012) quantify the relative importance of these factors empirically.

While concentration is important, the geospatial location of companies within regions is no less important. Strategic decisions, such as whether to locate in the center of a city or in the suburbs, the kind of office chosen by entrepreneurs and the link between these variables and office price, are key to guaranteeing business success. Factors such as information exchange facilities provided by technological breakthroughs and the rise in the price of properties in the center of major cities are behind the decisions of where to locate businesses.

In order to provide empirical evidence about this phenomenon, we analyze the relation between the price businesses are willing to pay for the location of service sector company headquarters and several price explanatory variables, among which geospatial location stands out. With this aim, we make use of the rental price per square meter data from 4,721 premises obtained in March, 2020 from the Idealista portal site. Geospatial location, measured in latitude and longitude, the surface area of the premises, number of bathrooms, and several features such as whether the premises has an exterior orientation or the availability of parking or an elevator, are used as explanatory variables.

To perform the analysis, the regression tree technique built upon a random forest algorithm was chosen. The versatility of these non-parametric methods allows us to study large databases where the relation between the variables in the model, even when qualitative, is featured by a strong nonlinear component. While not intended as an exhaustive review, the application of decision trees has been used to analyze spatial distributions in several studies, such as those by Fan, Ong and Koh (2006) to examine the price of dwellings, Nuruddin, Sitanggang and Yaakob (2014), who perform a spatial prediction of hotspots in peatlands, and Chelghoum and Zeitouni (2002) who analyze the risk of accident spatial distribution.

The main results of our study are as follows. First, we found by using the whole sample, that the variables which determine the price businesses are willing to pay for office rentals are those

that are established by geospatial location. Higher office rental prices are located in the city center. When the properties are farther away from the center in any direction, the price of rent decreases. However, the central axis of the Paseo de la Castellana remains the location where the highest rent is paid.

Secondly, in order to complement the analysis, a stratification of the data determined by office size was carried out. In particular, the data were classified according to the square meters of the offices, with small offices (up to 400 m^2), medium ones (between 400 m^2 and 800m^2) and large ones (over 800 m^2). Results point out that the concentration around the city center is not homogeneous for each of the three different kinds of offices.

The willingness to pay more in the case of larger offices is not as conditioned by the closeness to a geographic point. Some locations farther from the center are well valued because they have large rooms and are also located close to strategic transportation links. This result is in line with the findings by Holly and Stephens (1981) and Nijkamp and van Geenhuizen (2016), who, in addition, point out that less location-affected companies look for facilities that help improve work-life balance, and other services for employees in their workplace.

The degree of concentration of small offices around the city center is also unclear, probably due to the need the companies that rent those offices have to be located close to urban populations, although not necessarily in the center itself. Small, profitable businesses choose to be located in the center and pay high rent in order to be close to their clients. However, those less solvent also obtain profitability for their businesses, paying high rent far from the center, but always located in populated urban areas.

Medium-sized businesses are the ones with a more centrally located pattern, where higher rental prices are concentrated. The companies that rent medium-sized offices have enough financial profit to rent high priced offices in the center of the more exclusive large urban areas. However, these office spaces are not big enough for them to create their own services in the workplace when they move away from the center. These businesses only accept a location farther from the city center in exchange for lower rental prices.

The rest of the article is structured as follows. In Section 2, we describe the methods used to build the decision tree and the random forest method applied to control decision tree instability. In Section 3, the application to office rental prices in the municipality of Madrid is illustrated. Section 4 concludes.

2 Regression trees

In this section the regression tree technique is depicted. To facilitate a better understanding of the study, we start with the description of the CART algorithm. Then, we establish the basis for the implementation of the random forest algorithm. We recommend that readers already familiar with these techniques go directly to the empirical application.

2.1 CART algorithm

Although the first proposals for analysis with decision trees were present in the work by Morgan and Sonquist (1963), trees became popular with the CART algorithm (acronym for Classification and Regression Trees) proposed by Breiman et al. (1984).Venkata and Kiruthika (2015) make a comparison between CART and other popular construction tree methods, such as ID3, CART, C4.5, and See5/C5.0. These algorithms model a response variable from a set of p explanatory variables. Since in our case the response variable is the rental price of offices, which takes continuous numerical values and has an intrinsic order, regression trees are applied.[1]

Let's assume we have a sample of T observations, $\{(Y_1, X_1), \ldots, (Y_T, X_T)\}$, with Y_t is the observation of the dependent or response variable and $\mathcal{X}_t = (X_{1t}, \ldots, X_{pt})'$ the vector of observations of the p explanatory variables at t period, with $t = 1, \ldots, T$. The regression tree seeks to perform an orthogonal partition of the space of the possible values of the explanatory variables in a way that minimizes the loss function.

Since an evaluation of all the possible partitions that could be used is computationally unaffordable, the most common method is to perform recursive binary partitions by locally optimizing each of the partitions that give rise to two new sub-regions. At the beginning of the tree, the set of the possible values of the explanatory variables belongs to a single region. With the aim of creating the first partition, we will have to choose the division variable X_i and a division point or threshold c_i so that the whole is partitioned into two regions. The first region is created from the values of the whole, where the division variable is below the threshold $R^0 = \{\mathcal{X}_t | X_{it} \leq c_i\}$. The second one makes reference to the values of the whole for which the division variables are over the threshold $R^1 = \{\mathcal{X}_t | X_{it} > c_i\}$.

To select the partition variable and the division point, two additional components are needed. First, we should choose a forecast for the observations that belong to each of the regions, $\widehat{Y}(R^0)$ and $\widehat{Y}(R^1)$. In the case of regression trees, the mean of the dependent variable used is the

[1] When the response variable is qualitative, classification trees are applied.

observations that belong to each region

$$\widehat{Y}(R^d) = \frac{1}{N_d} \sum_{t \in R^d} Y_t, \tag{1}$$

where N_d is the number of observations in the region R^d and $d = \{0, 1\}$.

Secondly, we must define a loss function that allows us to choose the best partition among all the feasible partitions. In the context of regression trees, one of the most common loss functions is the quadratic

$$L\{Y, \widehat{Y}\} = \sum_{\mathcal{X}_t \in R^0} \left(Y_t - \widehat{Y}(R^0)\right)^2 + \sum_{\mathcal{X}_t \in R^1} \left(Y_t - \widehat{Y}(R^1)\right)^2, \tag{2}$$

which takes a small value when the observations from each region are close to the predictions made there. Finally, among all the explanatory variables, we select the variable and the division point that minimize the loss function.

Once the first partition is made, the procedure consists of repeating the previous process for each of the two resulting regions. The division variable and the threshold that minimize the loss function are found for each region, and the only region for which a smaller loss function is obtained is partitioned. Once the second partition is made, the procedure continues sequentially until some established criteria is reached. Sometimes the process is stopped when reducing the loss function resulting from a division below a threshold is not possible. In other cases, the process continues until partitioning is not possible without containing a minimum number of observations. To prevent over-fitting, pruning criteria consist of reaching a tree with the maximum number of end regions that are reduced to optimize some established loss function.[2]

In any case, the tree will result in a disjointed set of M terminal regions $\{R_1, \ldots, R_M\}$.Once the regression tree is built, it is very easy to make forecasts for the new regions. The forecast for a new observation will be the forecast for the final region to which it belongs. So, the forecast for the dependent variable Y_τ conditioned by the explanatory variables X_τ will be

$$\widehat{Y_\tau}(X_\tau) = \sum_{m=1}^{M} \widehat{Y}(R_m) I(X_\tau \in R_m), \tag{3}$$

where $\widehat{Y}(R_m)$ is the mean of the observations of the R_m region and $I(\cdot)$ is an indicator variable taking the value 1 for a true statement and 0 otherwise.

Two tools have been developed to analyze the results obtained with a decision tree. The first one, proposed in Breiman et al. (1984), consists of measuring the relative importance of X_i among the $(X_1, ..., X_p)$ available indicators. The measure is obtained based on the number of times this indicator has been used to make divisions in the tree, weighted by the loss function reduction caused by each of those divisions.

[2]Esposito et al.(1997) survey decision tree pruning techniques

The second regression tree analysis tool is the partial dependency graph, which measures the effect that different possible values of an explanatory variable X_i have on the dependent variable. Moreover, these graphs can be generalized by measuring the reaction of the dependent variable for any subset of feasible values of the explanatory variables. In this case, the graph will measure the interaction effects among the explanatory variables in predicting the dependent variable.

Decision trees belong to the supervised estimation method family. For this reason, to measure the predictive capacity of a decision tree, the sample has to be split into two subsamples: the training sample and the evaluation sample. The training subsample will be used to design the tree, and the evaluation subsample will be used to measure the predictive ability of the designed tree.

2.2 Random Forest algorithm

The simplicity of the tree construction process with the CART algorithm has a serious problem associated to it. Due to the fact that a small error made in the first splits of the tree will be magnified in the next branches and this leads to errors in the resulting prediction, instability results.

To solve this lack of robustness, several procedures based on the generation of intermediate trees that give rise to a final tree, obtained by averaging the intermediate trees, have been proposed. With the aim of reducing the above-mentioned instability, two proposals have been developed. The first one is to average complex trees generated in parallel through bootstrapping so that variance is reduced by controlling bias, such as bagging (Breiman, 1996) and random forest (Breiman, 2001) algorithms. The second proposal is to average simple trees generated sequentially to control bias by reducing variance, as in boosting (Freund and Schapire, 1997). In this paper we will follow the first strategy.

The first contribution in this context is the Bootstrap Aggregating (bagging) algorithm proposed by Breiman (1996). Applied to decision trees, the logic of the bagging algorithm is very simple. From the original sample B different samples are randomly generated, and with replacement by using the bootstrapping technique, $\{(Y_1^b, X_1^b), \ldots, (Y_T^b, X_T^b)\}_{b=1}^B$.[3] For each of these samples generated, a decision tree is estimated using the CART techniques described in the previous section so that each sample generated will result in a $\{R_1^b, \ldots, R_{M^b}^b\}$ partition. In the end, a final decision tree is constructed by averaging the results of the estimated trees in the generated samples. Therefore, the prediction of the bagging algorithm for the dependent variable

[3]Sub-samples can be generated without replacement (Subagging) although this option implies a computational cost with little impact on the predictive capacity

Y_τ conditioned to the observations of the explanatory variables X_τ will be

$$\widehat{Y}_\tau(X_\tau) = \frac{1}{B}\sum_{b=1}^{B}\widehat{Y}_\tau(X_\tau \in R_m^b), \tag{4}$$

where $\widehat{Y}_\tau(X_\tau \in R_m^b)$ is the prediction made for Y_τ conditioned to X_τ belonging to the R_m^b region in the b-th sample generated according to the expression (3).[4]

Bagging's predictive ability depends on the amount of trees used, although it usually stabilizes at the number of trees from which improvements are negligible. To select the number of trees used in the model, we usually start from a high number B^* and, through cross validation techniques, we determine the number $B < B^*$ from which the predictive capacity is stabilized. When the sample contains many observations, cross validation techniques can slow down the procedure. To speed up the search computationally, the out-of-bag method (OOB) is sometimes used, which consists of training each tree in 2/3 of the observations, leaving the rest to evaluate the predictive capacity.

Breiman (2001) showed that the ability of the bagging algorithm to reduce the resulting tree variance while maintaining the bias directly depends on the number of samples generated and, inversely, on the correlation between those samples. In practice, as the samples generated with bootstrapping are similar, since the trees are built with similar subsamples, the resulting trees will be highly correlated. It is also possible that an explanatory variable is much more important than the rest, and that all the trees use it as their initializer. This will also mean that the resulting trees will give rise to highly correlated predictions.

The most common solutions to correct the correlation between decision trees are based on the random forest algorithm initially proposed by Breiman (2001). The method consists of building the decision trees for each sample generated by bootstrapping from a subset of p^* variables, randomly chosen among the p explanatory variables in the sample. The number p^* of selected variables can be chosen with cross validation techniques or with OOB. In other cases, a number $p^* \leq p/3$ is simply chosen.

3 Empirical analysis: office rentals in Madrid

With the recent development of real estate portals, reliable sources of data that reflect potential real estate transactions have been made available. In addition, the detailed information about the characteristics of the advertised spaces and their geospatial location provide the possibility of carrying out detailed analyses of the price determinants of these spaces, specifically the possibility

[4] Bühlmann (2004) proposes other measures of central tendency, such as the median, which are more robust against outliers.

of spatial analysis.

In the empirical application that we perform in this section, we use a database of price per square meter for office rentals taken from the Idealista portal database, referencing 4,721 offices advertised on 23 March, 2020 . In addition to the price and latitudinal and longitudinal coordinates of the spaces, the database also includes information on characteristics of the advertised offices such as floor space, number of bathrooms, and different factors that refer to the availability of elevators, parking spaces, and exterior orientation.

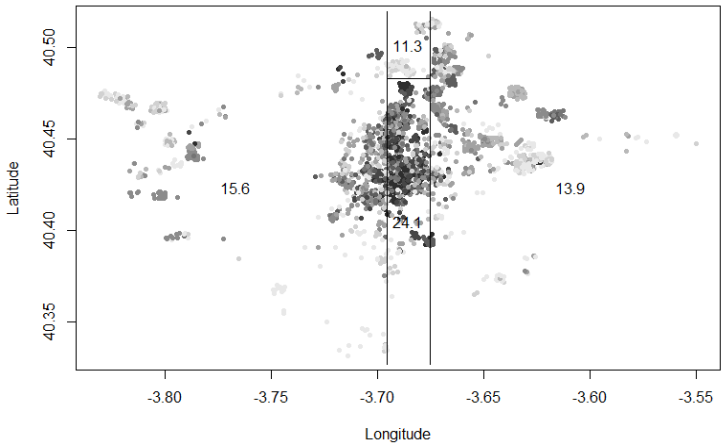
In the first part of this section, we will look at the determinants of office rental prices for the whole sample. However, we think that the access to an office with certain characteristics could depend on size. For this reason, in the second part of this section, we divide the sample for the analysis into large, medium-sized, and small offices.

3.1 Analysis of the whole sample

Figure 1 shows the location on the map of office prices per square meter in terms of longitude and latitude. To aid in understanding the information given, the points representing offices in the graphic are darker as the rental price increases. It can be observed that there is a marked geographic character in rental price distribution

The main characteristic of the map is the dense geographic concentration of offices for rent in Madrid; especially the highest-priced ones. Longitude plays an important role since the highest square meter prices are concentrated in the limited strip near the axis of the Paseo de la Castellana. However, office rental price distribution in this area is not homogenous. The most expensive offices are centered in the north, around the Four Towers; in the center, near Nuevos Ministerios; and between the Plaza de Colón and Banco de España. In the south, there is an area of high-priced rent near Mendez Álvaro, where Repsol has recently established its headquarters. The offices located farther north from the center tend to be much cheaper.

Figure 1: Price and location



In addition to the city center, there are small concentrations of high-priced rental areas, such as in Las Tablas and Sanchinarro. The area of Las Tablas is a new nucleus of large company headquarters (BBVA, Telefonica, Capgemini), as well as being near the main transportation links to the north. Another area where there is a group of expensive office rental prices is near the Madrid-Barajas Adolfo Suárez airport. Lastly, we can find high-priced office rentals in the west, in the area near the municipality of Pozuelo.

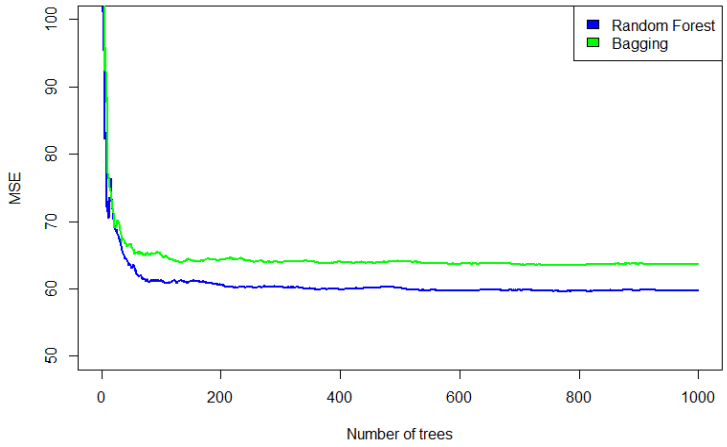
To understand how decision trees are built with CART algorithms, we a use a simple tree where rental price depends only on geographic coordinates to partition the map. To simplify the example, we will restrict the decision tree to give rise to a maximum of four regions, determined in a way that the quadratic loss function of the resulting partitions is reduced as much as possible. The most expensive offices are in the center, with an average price of 24.1 euros per square meter. This is followed by the western and eastern areas, with average prices of 15.6 euros and 13.9 euros per square meter, respectively. The cheapest area on the map is the area located to the north of the center, where the rental price for an office costs an average of 11.3 euros per square meter.

Obviously, geographic coordinates are not the only factor that influence office rental prices. To carry out a more thorough analysis of price determinants, we add the rest of the explicative variables. Moreover, to avoid the problem of instability, we will build the decision trees with the bagging and random forest algorithms, using three randomly chosen variables.

Figure 2 shows the evolution of the average loss function depending on the number of estimated trees in the samples generated by bootstrapping. For comparison, we add the result of boosting, although we will not use this technique because it provides worse results. As can be seen in the graphic, random forest produces lower loss function values. The gains of the algorithm are

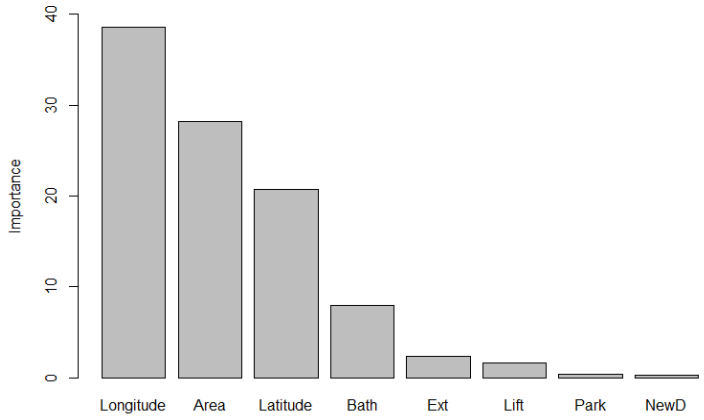
established beginning at 75 treess[5].

Figure 2: Comparison of algorithms



In Figure 3 the relative importance of the explicative variables in forming the resulting decision trees is analyzed. The geographic location of the office space clearly determines a large part of the square meter price in office rentals. The other variables are much less important.

Figure 3: Relative importance

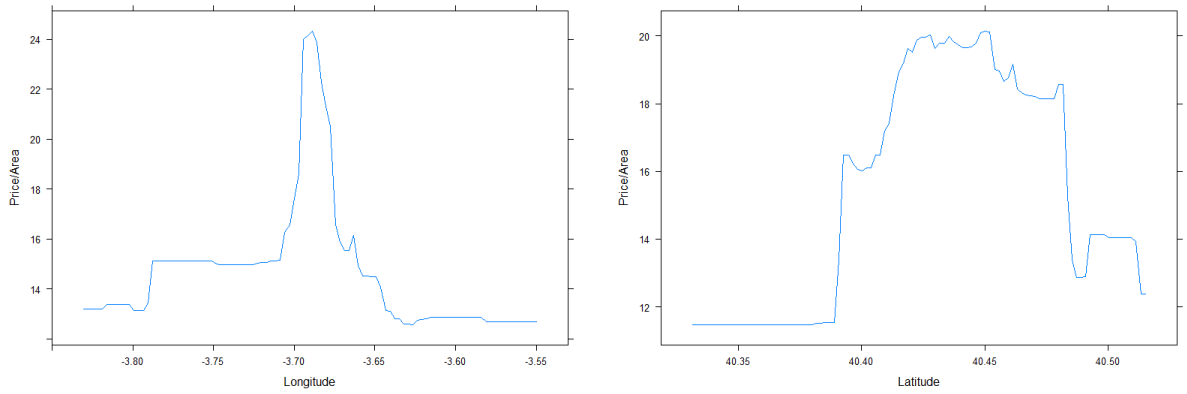


The relation between the explicative variables and office rental price is analyzed in the partial dependency graphics which are displayed in Figure 4. The effect of geographic location on the determination of rental prices is characterized by a marked nonlinear component. Beginning with longitude, when we move from the east toward the center, rental prices become more expensive up to the -3.7 longitude. In the center, which refers to the area up to the -3.68 longitude, prices are not influenced by latitude. From this point, prices are negatively related to longitude since

[5]This was also estimated using the boosting technique, obtaining values that were higher than the loss function

this implies moving away from the city center.

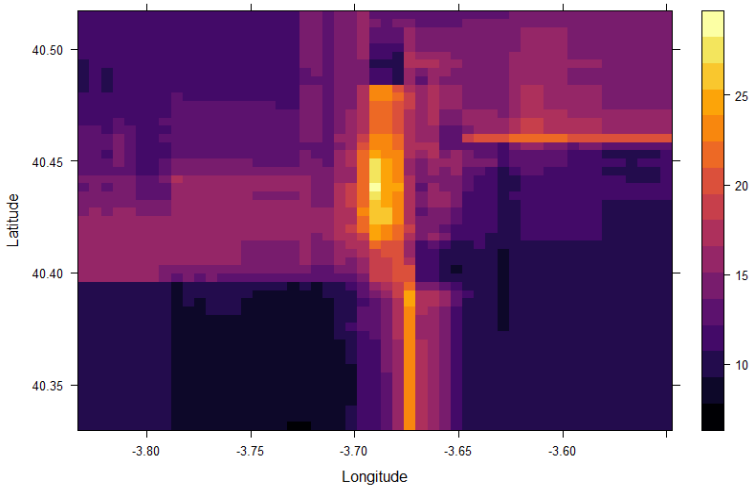
Figure 4: Partial dependence graphics



The relationship between the price of rent and latitude does not show a linearity similar to the one we found with longitude. Farther south than the 40.42 latitude, price is positivity affected by latitude. In the central part, between the 40.42 and 40.48 latitudes, rental prices do not seem to be associated to latitude. However, from this point, the relationship changes, and the price of office rentals decreases as latitude increases.

This close, nonlinear relationship between geographic location and rental prices is detected more intuitively in the interaction effects that appears in Figure 5. This type of graphic, made with random forest and all the variables, is more robust than that obtained with the partitions of the regression tree, and therefore, easier to use when analyzing spatial distribution. As the scale on the right shows, the spaces with the highest rent appear in the lightest color. A surprising factor is that office space rental prices do not display homogeneity in terms of geographic distribution. The center is the area that has the highest prices per square meter. However, in the area to the south of the center and in the northeast, rental prices are similar to those in the city center.

Figure 5: Interaction effects



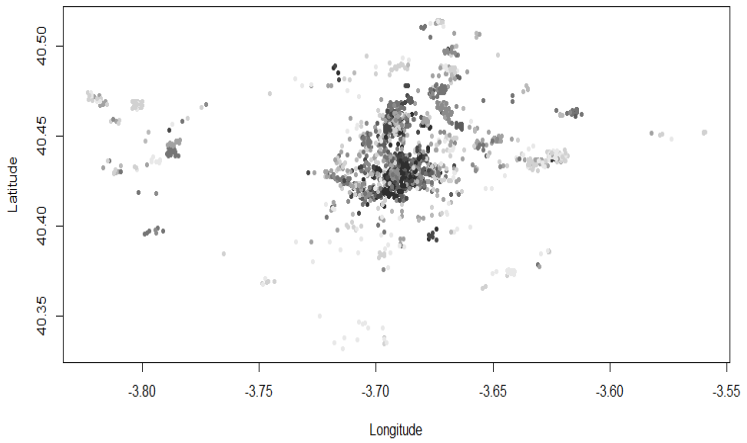
3.2 Stratified analysis

In this section, we will analyze the different situations that large companies face compared to small ones when deciding where to establish their offices, and what prices they are willing to pay to rent these offices. To carry out this analysis, we have divided the sample into three subsamples according to office size. In the first subsample, we include small offices of up to $400m^2$; in the second subsample, we include medium-sized offices of between $400m^2$ and $800m^2$; the last subsample includes offices of more than $800m^2$. These strata roughly correspond to observations of between 0 and 50, 50 and 80, and 80 and 100 percent, respectively, of the variable that measures office floor space.

3.2.1 Small offices

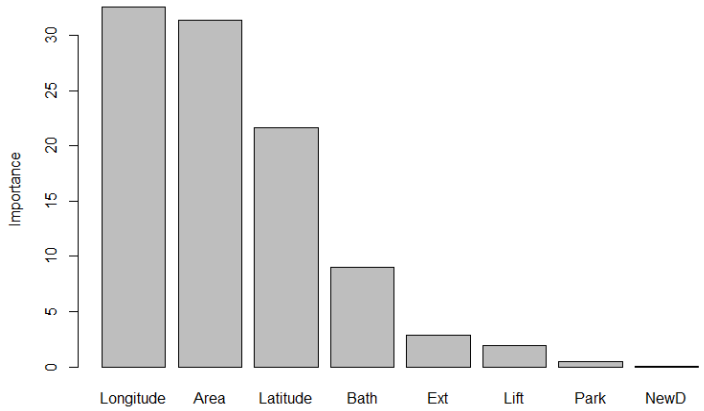
Figure 6 shows the geographic distribution of the offices for rent in the sample with a size of less than $400m^2$ and which make up about half of the total observations. The map is the most similar to the map of the whole sample, which appears in Figure 1 Although we can observe a concentration of expensive rental offices in the center, there are also small groups of relatively expensive rental prices on other parts of the map. This phenomenon can be explained by the need for small offices to be located near population centers, but not necessarily populations as dense as those in the center of Madrid.

Figure 6: Location and price for small offices



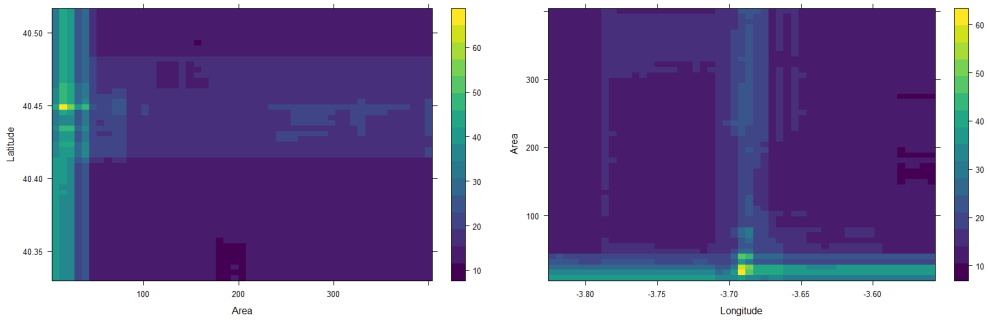
The relative importance of the variables that determine price appears in Figure 7. The variables that determine geographic location and surface area continue being the most important. In contrast to the result obtained for the whole sample, the surface area is now as important as geographic location. In addition, other variables such as bathrooms, elevators, or exterior orientation become much more important in determining the price of rent.

Figure 7: Relative importance for small offices



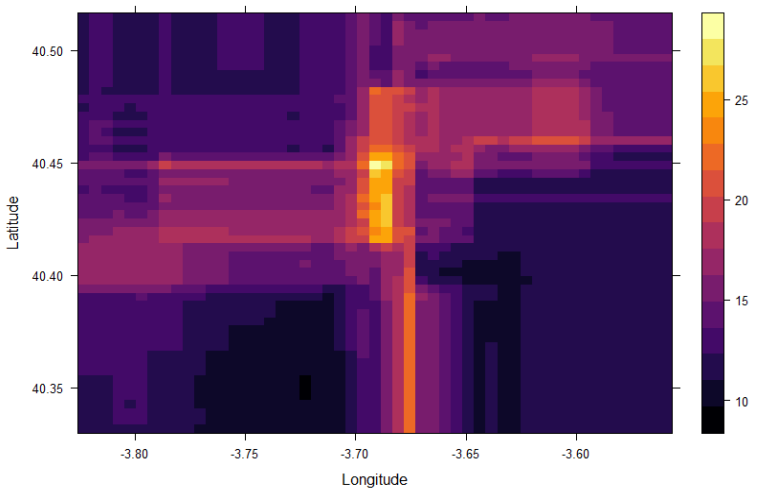
As can be seen in the interaction graphics of Figure 8, the distribution of the geographic location of small offices especially affects offices of less than $100m^2$. Neither the latitude (left graphic) nor the longitude (right graphic) significantly affect rental prices when they interact with offices with a surface area of less than $100m^2$. A possible exception could be offices with a longitude (and to a lesser extent, latitude), similar to that of the city center.

Figure 8: Interaction effects of latitude and longitude with surface area in small offices



The geographic distribution of small offices according to rental price can be better appreciated in Figure 9. As with the whole sample, we observe a willingness to pay higher prices per square meter to rent offices in the center of Madrid. However, there are areas appearing in a lighter color where high prices are paid in the entire area that are in the central and northeastern latitudes and longitudes of Madrid.

Figure 9: Interaction effects of location for small offices



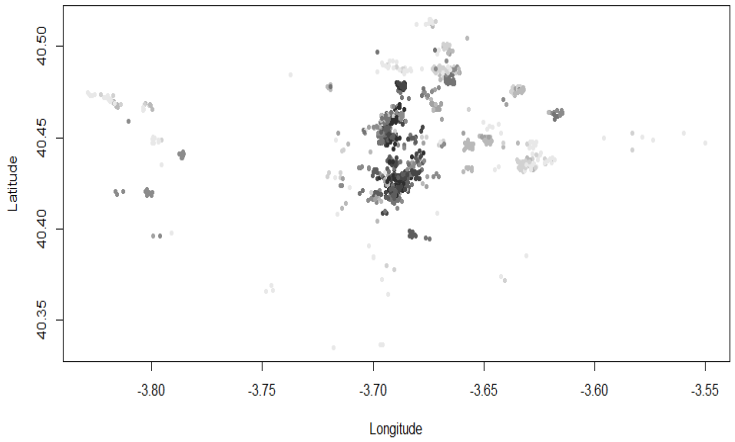
3.2.2 Medium-sized offices

In the case of medium-sized offices, which make up around 30% of the sample, the willingness to pay high rental prices is much more concentrated in the center. In Figure 10, which shows the geographic locations of medium-sized companies, we can see that the concentration of the darkest points, and therefore, the most expensive rent per square meter, is almost exclusively in the center of Madrid.

A possible explanation for this particular geographic distribution is that medium-sized companies only appear to be willing to consider locations outside the center of Madrid if they obtain

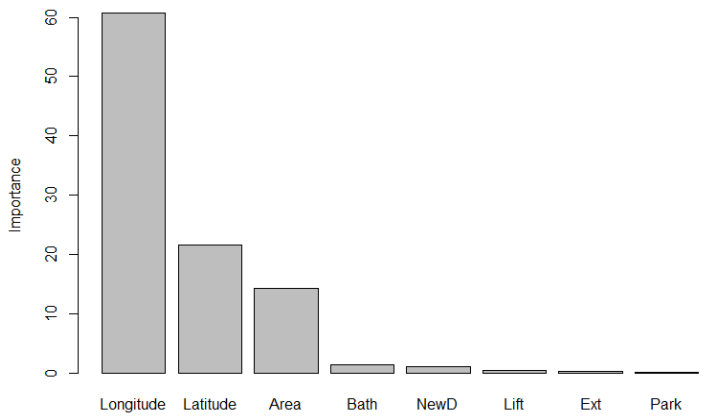
a substantial reduction in rental prices. As with small companies, medium-sized companies are concentrated close to population centers, which is where their clients are. However, it seems they need a minimum number of people to support their businesses. For this reason, they prefer to be in the center of Madrid, even though this implies paying higher prices to rent offices.

Figure 10: Location and price for medium-sized offices



In contrast to the case of small office spaces, the most important variables determining the price of rent for medium-sized office spaces are those having to do with their geographic location. As can be seen in Figure 11, which shows the relative importance of the explanatory variables, the surface area moves from second place to third place in terms of relative importance. By far, the most important variable determining rental price is longitude, indicating a preference to occupy an office space near the axis of the Paseo de la Castellana.

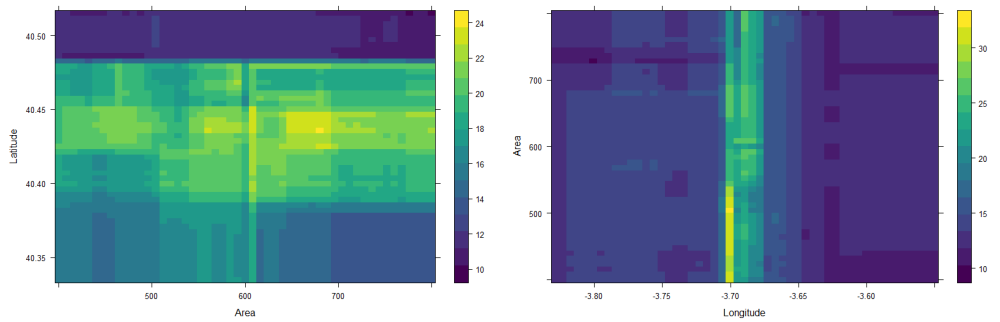
Figure 11: Relative importance for medium-sized offices



The interaction effects of latitude (left graphic) and surface area and longitude (right graphic)

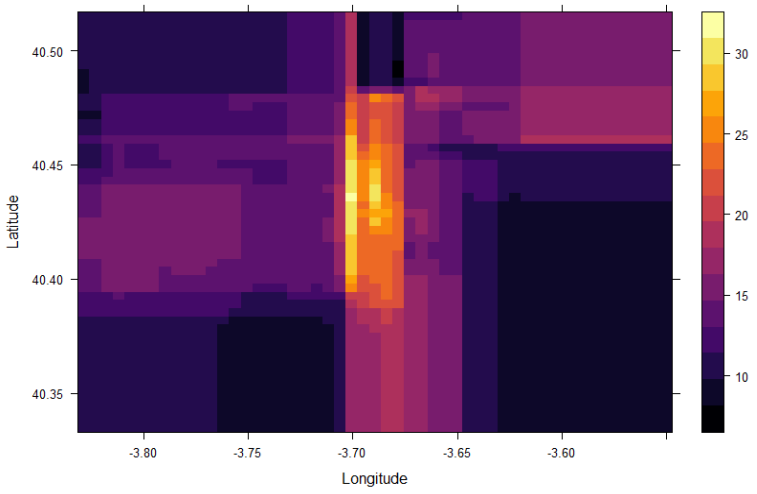
and surface area in determining rental prices is shown in Figure 12. The most expensive medium-sized offices are found at a much wider latitude band than in the case of smaller offices. Moreover, the effect of latitude on price appears throughout the distribution of medium-sized companies, in contrast to what occurs with small companies. Longitude appears to only affect price in the middle longitude, again concentrated around the axis of the Paseo de la Castellana.

Figure 12: Interaction effects of latitude and longitude with surface area in medium-sized offices



The decision tree model estimated with random forest is able to capture the nonlinear effect, which relates the geographic concentration of the highest rents for medium-sized companies shown in Figure 6. In order to detect this, Figure 13 shows the interaction effects of the latitudinal and longitudinal positions of medium-sized offices and their relation with rental prices. In the graphic, we can observe how the concentration of high prices appears in and around the center. The deviations with respect to the centralized location of medium-sized offices only occur when rental prices are significantly reduced.

Figure 13: Interaction effects of location for medium-sized offices



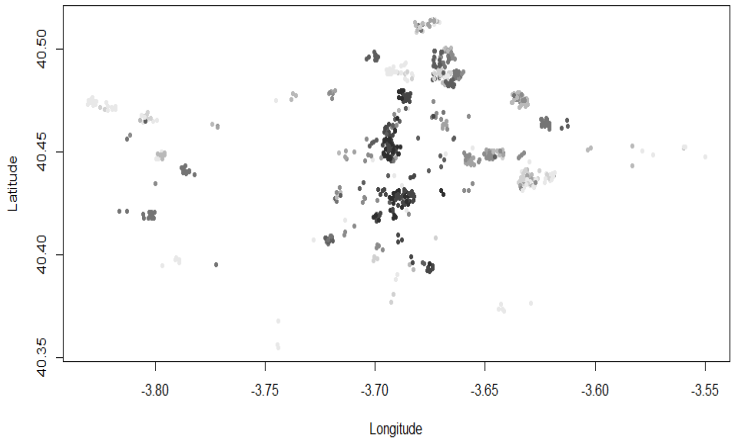
3.2.3 Large offices

The last group of companies included in the stratified analysis are the largest offices in the sample. As has been done with small and medium-sized offices, Figure 14 shows the geographic location of large office spaces, where the points representing the offices are darker as rental prices rise. The number of large offices makes up around 20% of the total, but the willingness to pay high prices for rent doesn't seem to be restricted to the center of Madrid.

As occurs with the other types of offices, the companies willing to pay the highest prices are concentrated around the axis of la Castellana. In this case, the concentration of the most expensive office spaces reaches the area of the Four Towers to the north, and Mendez Alvaro to the south. In addition to the center, there are small clusters of high-priced offices in other locations, such as in Las Tablas and Sanchinarro in the north. Specifically, in Las Tablas, there is a recently established nucleus of large companies (BBVA, Telefonica, Capgemini), and this area is also near principal northern transportation links.

Some areas far from the center are appreciated because they provide large surface space and are near strategic transportation links. Results obtained along these lines by Holly and Stephens (1981) and Nijkamp and van Geenhuizen (2016), indicate that the companies least affected by location seek infrastructures that facilitate work-life balance and other services for their employees located in the workplace.

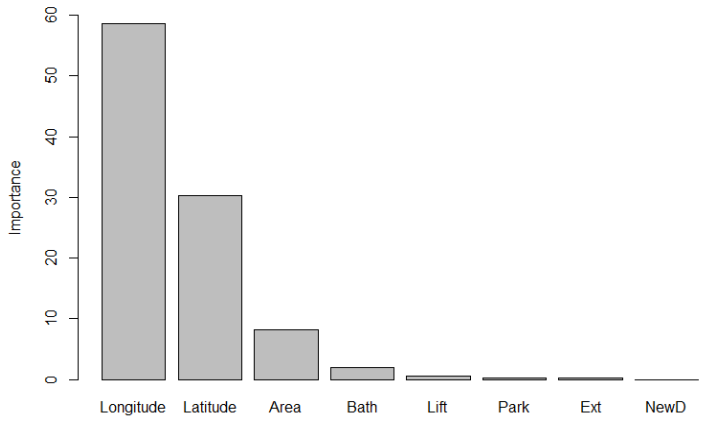
Figure 14: Location and price for large offices



In the case of large offices, surface area does not seem to be decisive when determining price, as Figure 15 shows. Companies that look for large offices need to find a good location. Therefore, the variables of longitude and latitude appear as the most important variables in determining the price of rent. In this case, both variables show a relative importance which is more similar

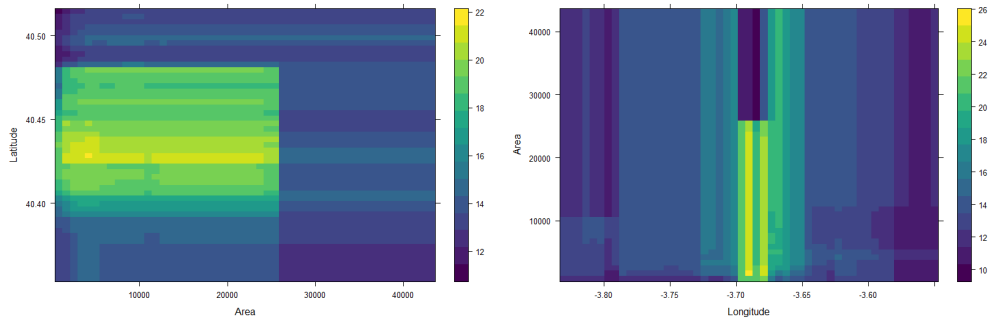
than in the case of medium-sized offices.

Figure 15: Relative importance for large offices



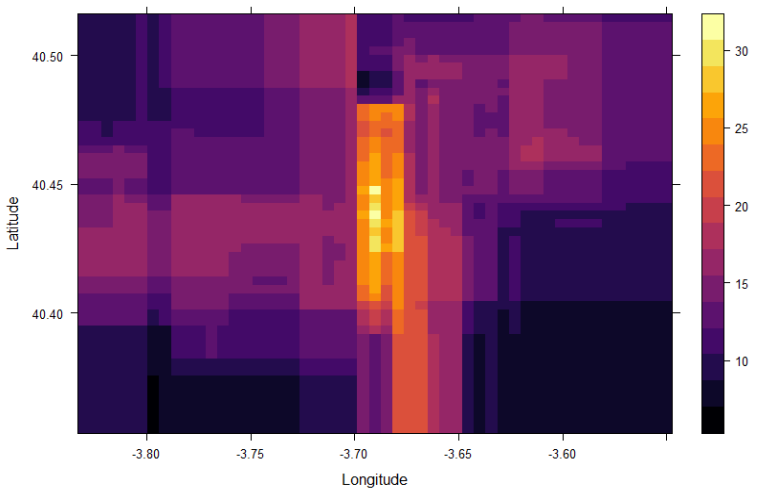
The interaction effect of latitude and longitude on surface area in determining large office prices is analyzed in Figure 16. As with medium-sized offices, and contrary to small offices, the effect of geographic location on rental prices is distributed homogeneously among large companies. However, this homogeneity seems to affect latitude more than longitude.

Figure 16: Interaction effects of latitude and longitude with surface area in large offices



Compared to the other two office groups, although companies show greater willingness to pay more for large offices in the center, there are many geographic areas which also have high-priced rent. This characteristic is illustrated in the interaction effects of location in determining rental prices shown in Figure 17. In this figure, the areas of expensive rent outside the city center are clearly differentiated.

Figure 17: Interaction effects of location for large medium-sized offices



4 Conclusions

The need for companies to locate their offices near their means of production is a phenomenon that has been thoroughly explored in the economic literature. In companies working in the service sector, it could be expected that population concentration in city centers would cause these companies to pay much higher office rental prices in order to be located in these centers. However, with the rise of new services and the generalized use of information technologies, this situation may not apply to all companies. To examine this phenomenon, we have analyzed the determinants that establish the price of office rentals in the municipality of Madrid, using decision trees estimated with random forest. The choice of this technique is because it allows us to capture the relationships among the explanatory variables and their interaction with the dependent variable with a marked nonlinear component.

Using a sample of 4,721 offices obtained from the real estate portal Idealista in 2020, we have analyzed the price per square meter determinants of office rentals. As explicative variables, the latitudes and longitudes marking the geographic position of the offices, the surface area of the offices, the number of bathrooms, and characteristics such as having an exterior orientation, parking space, or elevators, have been used. Our main result is that business are willing to pay high rental prices to locate their offices in the center of Madrid. However, we also found concentrations of high-priced offices far from the city center but located near urban populations or strategic axes.

To analyze the degree of homogeneity of this result, we have grouped the sample into small offices (less than $400m^2$), medium-sized offices (between $400m^2$ and $800m^2$) and large offices (more than $800m^2$). Our results show that medium-sized companies seek offices in the center

of Madrid, although this means paying a higher rental price. This situation also exists in small and large offices however, we find that there are also strategic locations for these small and large companies that lead them to pay elevated prices outside of the city center.

These results are in line with the literature on the decentralization of companies who seek larger office space since they are affected by their location in the city center. These companies look for bigger spaces in the context of companies who want to provide more services to their employees in the workplace, thereby facilitating work-life balance. These companies usually locate near strategic axes with good transportation links.

References

[1] Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24, 123-140.

[2] Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5-32.

[3] Breiman, L., Friedman, J., Stone, C. J. y Olshen, R. A. (1984). Classification and regression trees. New York: Chapman and Hall.

[4] Bühlmann, P. (2004). Bagging, boosting and ensemble methods. En J. Gentle, W. Härdle, y Y. Mori (Eds.) *Handbook of Computational Statistics: Concepts and Methods*. Berlin: Springer.

[5] Chelghoum, K., y Zeitouni, K. (2002). A decision tree for multi-layered spatial data *Advances in Spatial Data Handling*, Springer.

[6] Combes, P., Duranton, G., Gobillon, L., y Puga, D., (2012). The productivity advantages of large cities: distinguishing agglomeration from firm selection *Econometrica*, Volume 80,Issue 6, pp. 2543-2594.

[7] Duranton, G., y Puga, D., (2004). Micro-foundations of urban agglomeration economies. En J. Henderson y J. Thisse (eds), textitHandbook of Regional and Urban Economics. Elsevier, North Holland, New York.

[8] Ellison, G., Glaeser, E. L., y Kerr, W. R. (2010). What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns *American Economic Review*, 100, pp. 1195-1213.

[9] Esposito, F., Kay, J., Malerba, D., y Semeraro, G. (1997). A comparative analysis of methods for pruning decision trees *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 19, pp. 476-491.

[10] Fan, G., Koh, H. C., y Ong, S. E. (2006). Determinants of House Price: A Decision Tree Approach *Urban Studies*, 43(12), pp. 2301-2315.

[11] Freund, Y., y Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting *Journal of Computer and System Sciences*, 55(1), pp. 119-139.

[12] Gonzalez Val, R., y Marcén Pérez, M. (2018). Concentración empresarial y economías de aglomeración en Aragón *Consejo Economico y Social de Aragón*, Premio de investigación Ángela López Jiménez, 2017.

[13] Holly, B.P., y Stephens, J.D. (1981). City system behaviour and corporate influence: the headquarters location of US industrial firms, 1955-75 *Urban Studies*, 18, pp. 285-300.

[14] Marshall, A. (1890). Principles of Economics *London: Macmillan.*

[15] Melitz, M. (2003). The impact of tade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6), pp. 1695–1725.

[16] Melitz, M., y Ottaviano, G. (2003) Market size, trade and productivity. *Review of Economic Studies*, 75 (1), 295–316.

[17] Morgan, J. y Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Jounal of the American Statistical Assocciation*, 58, pp. 415–434.

[18] Nijkamp, P., y van Geenhuizen, M. (2007). Cities and footlooseness: in search of place-bound companies and effective location policies *Environment and Planning C: Goverment and Policy*, 25(5), pp. 692-708.

[19] Nuruddin, A., Sitanggang, I., y Yaakob, R.(2014). A decision tree based on spatial relationships for predicting hotspots in peatlands *TELKOMNIKA Telecommunication Computing Electronics and Control*, 12, pp. 511-518.

[20] Puga, D. (2010). The magnitude and causes of agglomeration economies *Journal of Regional Science*, 50(1), pp. 203-219.

[21] Smith, A. (1776). An inquiry into the nature and causes of the wealth of nations. London: Printed for W. Strahan and T. Cadell.

[22] Venkata, S., y Kiruthika P. An overview of classification algorithm in data mining. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(12), pp. 255-257.