Combining Quality Indexes in the Retail Location Problem Using Generalized Linear Models

Virginia Ahedo, José Ignacio Santos, and José Manuel Galán(✉)

Departamento de Ingeniería de Organización, Escuela Politécnica Superior, Universidad de Burgos, Ed. A1, Avda. Cantabria s/n, 09006 Burgos, Spain
{vahedo,jisantos,jmgalan}@ubu.es

Abstract. The most important strategic decision in retailing is location. The process of selecting a proper place is a complex and multidimensional problem. A relevant factor that must be taken into account in the decision is the existence of an appropriate commercial ecosystem for the type of business to be located. There are different network-based quality indices to quantify the fitness of each location. In this paper, we show that the combined use of all the primary quality indices through generalized linear models and the aggregation of the information through consensus techniques allow improving the assessment of the different locations.

Keywords: Complex networks · Retail location problem · Prediction · Knowledge transfer · Classification · Pattern recognition · Generalized linear models

1 Introduction

Location is probably the most relevant strategic decision in retailing. The difficulty of completely imitating this aspect can be a critical competitive advantage [1].

The retail location problem is complex and multidimensional. Therefore, it is common practice to assess the different and varied factors that influence the decision and then to evaluate the available choices using multi-criteria decision techniques [2, 3].

An important dimension in the decision is the adequacy of the commercial ecosystem of the neighborhood for the type of activity to be located. Different network-based techniques analyze the location patterns of business categories (bakeries, bars, restaurants, etc.) to identify the level of attraction and repulsion between them [4, 5]. A set of primary quality indexes has been proposed to condense this information, each based on various assumptions about the structure of the commercial network and/or quantification of business patterns. Previous work suggests that the combination of all available metrics, given their complementary perspectives, may be the most interesting approach [6].

In this paper, we analyze the effect of using all the primary quality indices as features of Generalized Linear Models (GLM) to obtain a classifier capable of predicting the

F. P. García Márquez et al. (Eds.): CIO 2022, LNDECT 160, pp. 47–52, 2023.
https://doi.org/10.1007/978-3-031-27915-7_9

commercial category at each location. Classifiers with high performance and predictive capacity can be used as tools to comparatively assess the suitability of each location alternative and to enhance existing location recommender systems.

2 Generalized Linear Models

Generalized Linear Models are a family of classification and regression techniques that generalize traditional linear regression models in two ways: (i) for target variables that follow some exponential distribution and (ii) by allowing the variance to be dependent on the estimated value [7]. Given the nature of the location problem, the response variable has been modeled as a multinomial distribution.

The two most important hyperparameters of this family of models are *alpha* and *lambda*. Alpha is a regularization parameter that determines the type of regularization applied to the model and varies within the range [0,1]. When it is 0, the model is known as Ridge regression (L2-regularization), while when it is 1, the model is known as Lasso (L1-regularization); for intermediate values, where both types of regularization are combined, models are known as elastic nets [8, 9]. The particular regularization value set by the model with respect to each family is given by the lambda parameter and is obtained by cross-validation. L2-regularization allows dealing with problems of high correlation between variables, while L1-regularization allows obtaining a more parsimonious and sparser model through the selection of variables.

3 Computational Experiments and Results

The computational experiments have been conducted on data from the nine provincial capitals of Castile and Leon (Spain). The dataset used is publicly available as open data [10]. The commercial information in such dataset includes the business typology of the stores listed in the Yellow Pages, their geo-location extracted from MapQuest, Open Street Map, and Google Maps, and the proximity networks (for a radius of 100 m).

The business categories are classified according to the North American Industry Classification for Small business (NAICS) [11]. This type of classification has been used in previous research on retailing [4–6, 12, 13]. The NAICS establishes 68 different business categories.

The performance metric used to compare the different algorithms is the Mean Reciprocal Rank (MRR) (see Eq. (1)) [14].

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i} \tag{1}$$

The MRR is a statistical metric to evaluate the quality of the response based on the position in which the correct response is placed. Typical performance measures for classification such as accuracy and/or other metrics based on the confusion matrix are not adequate for the problem. The reason is that, although the empirical assignment of a category to a certain location is evidence of the suitability of the location for the category, this does not mean that the location is not suitable for alternative categories. Given this, a high-performing algorithm is one that scores the actual empirical category high in the ranking of possible categories for the location under scrutiny, but not necessarily in the first position.

The training and evaluation process has been performed using data from eight cities as the training set and the data from the remaining city as the test set. This process has been rotated to obtain nine evaluations of each algorithm (the results for each city are shown in Fig. 1, where the color of the dots represents the MRR obtained with each algorithm for the same city). To aggregate the information of the eight cities in the training datasets, the consensus networks of relationships methodology [4] was used without applying any threshold. The lambda and alpha hyperparameters of the GLMs have been optimized by conducting 5-fold cross-validation on the training data. The results have been compared with those obtained using the six primary quality indices based on networks proposed in the scientific literature [4–6, 12, 13]. These are Quality Jensen (QJ), Quality Permutation (QP), Quality Rewiring (QR), Quality Jensen Raw (QJR), Quality Permutation Raw (QPR), and Quality Rewiring Raw (QRR).

The six quality measures are based on Eqs. (2) and (3), where X represents Jensen, Permutation or Rewiring, the three alternative ways to obtain the interaction matrices a_{ij} between the different business typologies. N represents the total number of different categories, $nei_{ij}(x, y)$ denotes the number of neighbor stores from category j around the geographical point (x, y) (assuming that (x, y) belongs to category i given a considered interaction radius), $\overline{nei_{ij}}$ indicates the average number of neighbors of category j that the stores of type i empirically have. Equation (2) states how the candidate location fits the empirical proportion, while Eq. (3) aggregates—duly weighted—the attraction and repulsion relations in the vicinity of the considered point.

$$Q_{X_i}(x, y) \equiv \sum_{j=1}^{N} a_{ij}\left(nei_{ij}(x, y) - \overline{nei_{ij}}\right) \tag{2}$$

$$Q_{X-RAW_i}(x, y) \equiv \sum_{j=1}^{N} a_{ij}\left(nei_{ij}(x, y)\right) \tag{3}$$

The results of the analyses are shown graphically in Fig. 1. The GLMs successfully combine the metrics and improve the results of any of the primary quality indices used in isolation. These results support the notion of complementarity of the different metrics and the possibility of aggregating them effectively. Better predictors allow for improved assessment of the quality of commercial ecosystems and more successful evaluation of the suitability of different locations.

However, although the results seem straightforward, it is relevant to establish the degree of confidence in the conclusions, especially given that the number of cities is limited since the data collection and analysis process is computationally expensive. For this purpose, we have used the Bayesian Signed-Ranked test [15, 16]. This test allows to establish the probability, given the available evidence, that one algorithm is better than the other or—by setting limits that determine a region of practical equivalence (rope)—that both algorithms can be considered equivalent. The results of this analysis comparing the best quality index found empirically in our data (QJR) with GLMs are shown in Fig. 2. For 10000 Monte Carlo samples, we found that the probability that GLMs are better than the best of the primary quality indices used in isolation (QJR) is close to 0.75.

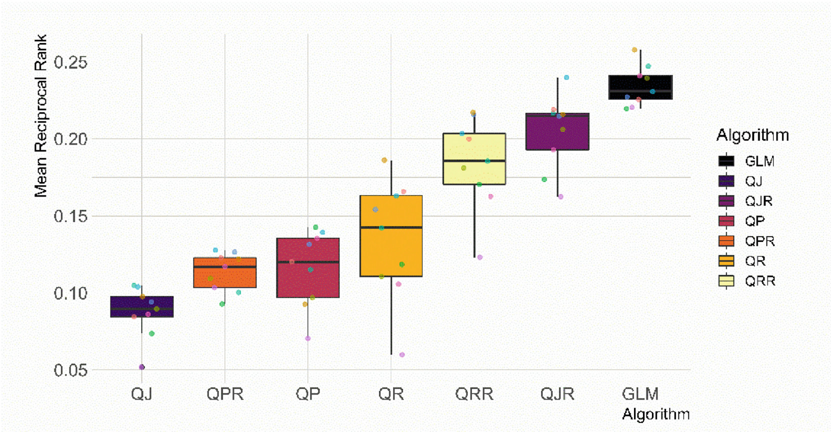

Fig. 1. Comparison of the predictive performance of the primary quality indices in isolation and that of the generalized linear models. The color of the dots represents the MRR results for each city.

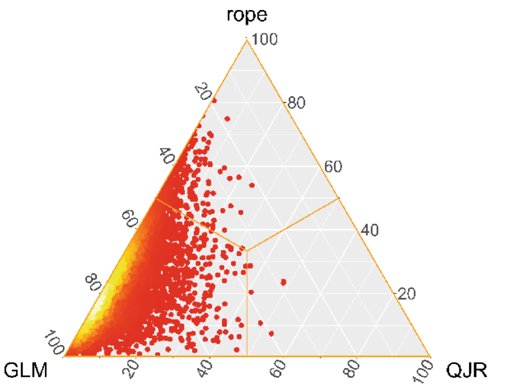

Fig. 2. Results of the Bayesian signed-rank test for comparison between algorithms. The results compare the GLMs with the best of the empirically found quality indices, the QJR. Given the evidence from the data, the probability that the GLM is better is 0.734, the probability that the QJR is better is 0.033, while the region of practical equivalence (rope) for a range [−0.01, 0.01] is 0.233.

4 Conclusions

The location problem in retailing is a very relevant problem in strategical terms, but at the same time, it is complex and multidimensional. An influential dimension in the decision is the suitability of the commercial ecosystem in the candidate neighborhood. Different quality indices based on network theory attempt to quantify this adequacy based on various assumptions. In this work, we have shown that using several sources of information from different cities, aggregated through consensus techniques, and the combination of all quality indices using generalized linear models improve the predictive performance and, consequently, the assessment of potential locations. These results suggest that the combined use of the primary quality indices aggregated by means of supervised learning techniques is a better performing approach than using them in isolation.

Acknowledgements. The authors acknowledge financial support from the Spanish Ministry of Science, Innovation and Universities (Excellence Network RED2018-102518-T), the Spanish State Research Agency (PID2020-118906GB-I00/AEI/10.13039/501100011033) and the Fundación Bancaria Caixa D. Estalvis I Pensions de Barcelona, La Caixa (2020/00062/001). In addition, we acknowledge support from the Santander Supercomputación group (University of Cantabria), that provided access to the Altamira Supercomputer—located at the Institute of Physics of Cantabria (IFCA-CSIC) and member of the Spanish Supercomputing Network—to perform the different simulations/analyses.

References

1. Zentes, J., Morschett, D., Schramm-Klein, H.: Strategic Retail Management. Gabler Verlag, Wiesbaden (2012). https://doi.org/10.1007/978-3-8349-6740-4
2. Shaikh, S.A., Memon, M.A., Prokop, M., Kim, K.S.: An AHP/TOPSIS-based approach for an optimal site selection of a commercial opening utilizing geospatial data. In: Proceedings - 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020, pp. 295–302 (2020). https://doi.org/10.1109/BigComp48618.2020.00-58
3. Çoban, V.: Solar energy plant project selection with AHP decision-making method based on hesitant fuzzy linguistic evaluation. Complex Intell. Syst. **6**(3), 507–529 (2020). https://doi.org/10.1007/s40747-020-00152-5
4. Sánchez-Saiz, R.M., Ahedo, V., Santos, J.I., Gómez, S., Galán, J.M.: Identification of robust retailing location patterns with complex network approaches. Complex Intell. Syst. **8**(1), 83–106 (2021). https://doi.org/10.1007/s40747-021-00335-8
5. Jensen, P.: Network-based predictions of retail store commercial categories and optimal locations. Phys. Rev. E **74**, 035101 (2006). https://doi.org/10.1103/PhysRevE.74.035101
6. Ahedo, V., Santos, J.I., Galan, J.M.: Knowledge transfer in commercial feature extraction for the retail store location problem. IEEE Access **9**, 132967–132979 (2021). https://doi.org/10.1109/ACCESS.2021.3115712
7. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2009). https://doi.org/10.1007/b94608
8. Tibshirani, R.: Regression selection and shrinkage via the lasso. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **58**, 267–288 (1996). https://doi.org/10.2307/2346178
9. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. Ser. B (Stat. Methodol.) **67**, 301–320 (2005). https://doi.org/10.1111/j.1467-9868.2005.00503.x
10. Sánchez-Saiz, R.M., Ahedo, V., Santos, J.I., Gómez, S., Galán, J.M.: Dataset of the retailing location networks in the cities of Castile-Leon, Madrid and Barcelona (2021). http://hdl.handle.net/10259/5585. https://doi.org/10.36443/10259/5585
11. Statistics Canada Government of Canada: NAICS 2012 – 44-45 - Retail trade (2011). http://www23.statcan.gc.ca/imdb/p3VD.pl?Function=getVD&TVD=118464&CVD=118465&CPV=44-45&CST=01012012&CLV=1&MLV=5
12. Gómez, S., Jensen, P., Arenas, A.: Analysis of community structure in networks of correlated data. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **80**, 16114 (2009). https://doi.org/10.1103/PhysRevE.80.016114

13. Jensen, P.: Analyzing the localization of retail stores with complex systems tools. In: Adams, N.M., Robardet, C., Siebes, A., Boulicaut, J.-F. (eds.) IDA 2009. LNCS, vol. 5772, pp. 10–20. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03915-7_2
14. Craswell, N.: Mean reciprocal rank. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems, pp. 1703–1703. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-39940-9_488
15. Carrasco, J., García, S., Rueda, M.M., Das, S., Herrera, F.: Recent trends in the use of statistical tests for comparing swarm and evolutionary computing algorithms: practical guidelines and a critical review. Swarm Evol. Comput. **54**, 100665 (2020). https://doi.org/10.1016/j.swevo.2020.100665
16. Benavoli, A., Corani, G., Demšar, J., Zaffalon, M.: Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. J. Mach. Learn. Res. **18**, 1–36 (2017)