Spatial data mining for retail sales forecasting

Maike Krause-Traudes[1], Simon Scheider[1], Stefan Rüping[1], Harald Meßner[2]

[1] Fraunhofer-Institut Intelligente Analyse- und Informationssysteme (IAIS), Sankt Augustin, Germany

[2] Rewe Group Austria, Wien, Austria

Abstract. This paper presents a use case of spatial data mining for aggregate sales forecasting in retail location planning. In particular, the data mining technique Support Vector Regression (SVR) is used to design a regression model to predict probable turnovers for potential outlet-sites of a big European food retailing company. The forecast of potential sites is based on sales data on shop level for existing stores and a broad variety of spatially aggregated geographical, socio-demographical and economical features describing the trading area and competitor characteristics. The model building process was guided by a-priori expert knowledge and by analytic knowledge which was discovered during the data mining process itself. To assess the performance of this SVR-model, it is tested against the traditional state-of-the-art gravitational Huff-model. Findings not only reveal that the spatial data mining model highly outperforms the traditional modeling approach with regard to prediction accuracy, but also that knowledge which is hidden in data and became discovered by the data mining process is of particular importance for constructing a valid and precise prediction model.

Keywords: Spatial Data Mining, Support Vector Regression (SVR), store location planning, aggregate retail sales forecasting, Business Geographics, Knowledge Discovery in Databases (KDD)

1. INTRODUCTION

Choosing the right site is crucial for the success of every retailing company because location decisions determine the external market conditions and the internal scope for action. In operational practice return on investment is the most important decision criterion for occupying a new site. From this point of view, location decisions are microeconomic investment problems. Consequently, the prediction of turnover for a potential new site becomes essential for every location planning decision. But the sales forecast still poses one of the greatest challenges of retail location planning of today.

But how precise can potential sales be predicted?

To answer this question this article presents an empirical comparison of two different retail sales forecasting models. The results of a Support Vector Regression model based on spatial data mining on the one hand are compared to a forecast derived by the classical gravitational Huff-model on the other hand. The Huff-model is adopted for comparison because it is of big theoretical importance and surveys show, that it is still widely used in retailing practice (cf. Hernandez, 2000) today**.** To compare their forecasting quality, both approaches are applied to sales data of about 870 outlets of a big European food retailing company from November 2004 to November 2005.

The paper is organized as follows: after a short delineation of the state-of-the-art in the field of sales forecasting, the next section introduces the Huff-model. Section 3 gives a short introduction to Support Vector Regression; section 4 contains a short description of the use case. For reasons of confidentiality, no company-specific details can be given here but an outline of the used geographical data. Section 5 describes the adopted forecasting models and statistical error measures. Section 6 and 7 contain the discussion of the modeling results and the conclusion.

2. HUFF'S MODEL OF RETAIL GRAVITATION

The initial development of systematic models for retail sales forecasting took place in the US-American marketing geography and started after World War II. During the 1960ies three basic approaches were developed which became the cornerstone for every retail sale forecasting model: the analogue approach developed by W. Applebaum's (1966), Huff's Model of Retail Gravitation (1964) and the deployment of multivariate regression for retailing questions (Megee, 1968). The analogue approach meets forecasts by assigning the turnovers of existing analogue stores (similar in terms of outlet and trade area characteristics) to a potential new site. The regression approach on the other hand builds up on the idea that a dependent variable - mostly the sales of a potential site - can linearly estimated by a series of explanatory variables describing the outlet and trade area characteristics (Craig, 1984). The concept behind the Huff-model is that the attractiveness of an outlet "gravitates" sales while the distance between offering and demand location deters the consumer from frequenting a specific location. A more detailed description of the model of retail gravitation follows in the next paragraph.

Since the 1990ies a revival of investigation into sales forecasting took place. Technological progress allowed the companies to collect, store and analyze vast amounts of data. Soon the idea grew that knowledge hidden in this operative datasets could improve retail location planning in general and sales forecasting in particular. The key to revelation of this hidden knowledge in vast data sets lies in the field of data mining. The authors believe that algorithms of data mining and artificial intelligence can be used to improve the existing approaches. For this reason the initial question shall be refined:

How much better are modern data mining approaches compared to traditional methods in prediction sales for potential new sites? Do improvements in forecasting accuracy, which could be expected by more complex modeling approaches, really justify the considerably higher model development input?

The group of gravitational models derives its name from the analogy to Newton's physical law of gravitation. Reilly was the first to assign the socio-gravitational considerations to the shopping behaviour of consumers with his "Law of Retail Gravitation" (Reilly, 1939). It states that the proportion of retail sales of two cities attracting customers from a third city is proportional to their respective attractiveness and inversely proportional to the ratio of weighted distances between offer and demand locations (Fittkau, 2004). Because the relationship between sales on the one hand and attraction and distance on the other hand is not necessarily linear, he introduced exponents for attraction and distance (alpha α and lambda λ). These exponents express the consumer's perception of attraction of and distance to the outlets. They are different not only between commodity groups but also between different geographical regions. Therefore, α and λ have to be estimated anew for every investigation area (Kotschedoff, 1976).

In 1964, Huff extended the bipolar model of Reilly in which consumers may only choose between two offering locations to a multidirectional model. In this model consumers are drawn to a certain offering location depending on its attractiveness and the economical distance to it while the perceived attraction declines with increasing distance. Huff also emphasized that consumers normally can choose between more than two alternate offering locations and take additional factors into account besides distance. That makes his model probabilistic. The formula to calculate these patronage probabilities – the Huff-model – can be formally expressed as:

$$p_{ij} = \frac{A^{\alpha}{}_{j} * d_{ij}{}^{-\lambda}}{\sum_{j=1}^{n} A^{\alpha}{}_{j} * d_{ij}{}^{-\lambda}} \qquad (1)$$

where

p_{ij} be the probability that a consumers from demand region i patronizes facility j

A_j be the attractiveness of facility j for j = 1,…., n

d_{ij} be the distance between demand region i and facility j for i=1, …, n and j=1, …, m

α be the parameter of attractiveness (normally equal to 1)

λ be the friction of distance

By multiplying the probability value p_{ij} with the absolute buying power Kk_i of the respective subarea, it is possible to get a quantification of the capital flow from subarea i to location j. By summing up the different capital flows directed to the locations, the probable turnover for a location can be estimated:

$$U_j = \sum_{i=1}^{n} p_{ij} * Kk_i \tag{2}$$

for j = 1,2,….,m outlets and i = 1….n sub areas

One of the biggest problems in making a Huff forecast of expected sales is the determination of lambda. It can either be assessed empirically or by statistical means like the Ordinary Least Squares approach (cf. Nakanishi & Cooper, 1974).

3. SALES FORECASTING AND SPATIAL DATA MINING

Data mining (DM) is commonly defined as extraction of implicit, yet unknown and potentially useful information from data by using methods from statistics, artificial intelligence, machine learning and pattern recognition. Today, DM is seen more and more as one step of a systematic and iterative knowledge discovery process, in which automated pattern recognition methods are combined with the analyst's expert knowledge. This process is known as Knowledge Discovery in Databases (KDD). Spatial data mining (SDM) is a special form of DM. The main difference to general DM in relational databases is that in case of SDM the relations considered are spatial relations such as topology or distance (Rinzivillo, 2008). SDM is of special value for business geographics, traffic control or environmental studies. The DM algorithm which was used in this case study is from the field of Support Vector (SV) algorithms. They were originally developed to solve classification problems and have been extended to deal with regression problems as well (Vapnik, 1997). SV methods are widely used today in a broad variety of scientific fields, e.g. for gene sequence analysis. As far as we know, Support Vector Regression (SVR) has not yet been used to solve the regression problem in geographical sales forecasting. We used SVR in this case because the data preprocessing revealed linear relationships between sales and explanatory variables which can be used by a multivariate regression prediction model.

Principles of SVR

The goal in supervised learning is to choose a function from the hypothesis space which predicts a target attribute most accurately with regard to some defined error measure. The target space of this function can either be nominal – as in the case of classification - or ratio scaled. SV algorithms are specially designed to minimize the expected classification error by minimizing both the empirical error and some measure of complexity. SVR uses more or less the same principles as the Support

Vector Classification. Consider the example case of a binary classification of two-dimensional data (represented by dots in figure 1). An unknown instance shall be assigned to one of the two classes. This task is operationalised as finding a line which separates these classes in an optimal way. In most cases, there are a lot of possible solutions to separate the classes, as shown in figure 1.

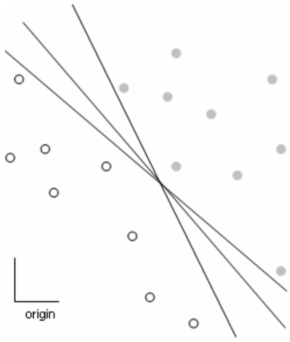

Figure 1 ***Possible linear classifiers***

The SV-approach searches for the linear classifier which separates the positive from the negative instances of the training set by maximising the margins. The margin is the distance between the separating line and the nearest data points (the Support Vectors). This linear classifier is called the Optimal Separating Hyperplane. The hyperplane H can be described by the orthogonal vector w and the offset b, such that $H = \{x|wx+b = 0\}$. However, in many practical cases, the instances are not linearly separable (like shown for the two-dimensional case on the left side of figure 2). In that case kernel functions can be used. The idea behind the kernel is to transform the input space containing the training instances into a new, higher-dimensional feature space, in which it becomes possible to separate the data.

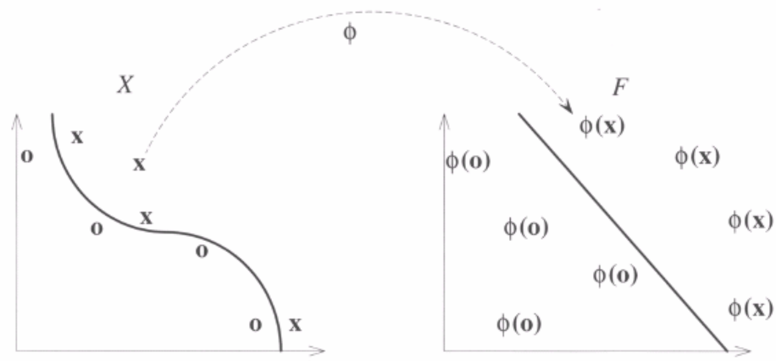

Figure 2 Kernel-transformation (Christianini, 2000)

Now consider the problem of approximating a set of data with numerical target values by a linear function

$$f(x) = \langle w, x \rangle + b \,.$$

To solve such a regression problem, the SV approach can be adapted, when instead of the classification error a suitable regression error is used. Again, the goal is to minimize the error and maximize the margin, which now corresponds to the task of finding the smoothest function that matches the data with an error of at most ε. A soft margin determines a penalty when the predicted value differs from the actual value more than the quantity ε. This penalty is expressed by slack variables ξ_i and ξ_i^* for over- and underpredicting (see figure 3).

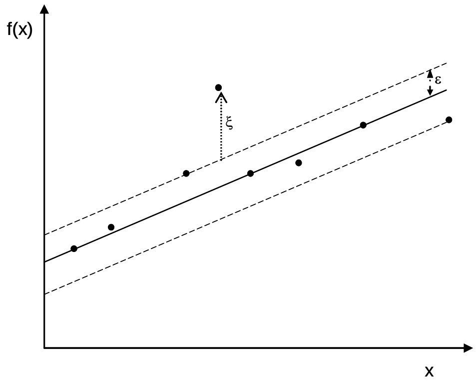

Figure 3 *ε -intensive loss function and slack variables*

In case of SVR the optimal regression function is given by

$$J = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{3}$$

where C is a given value. New values are computed with (Gunn, 1998):

$$y(x) = \sum_{i=1}^{n} \beta_i x_i * x + b\ . \tag{4}$$

Hence, applying SVR for sales forecasting means to provide a set of existing outlets with their sales and describing attributes like floor space, number of parking lots, distance to the next competitor etc. and infer from these training examples the relationship between sales and the describing attributes. This relationship is expressed by the regression coefficients in formula (4). For a potential new site, the expected sales value y(x) can be calculated by inserting the location specific values of the explanatory attributes into the regression equation.

4. CASE STUDY

The case study was based on data from the spatial data warehouse of a big European food retailing company. The data available for empirical comparison included polygonal and point map data, attribute data of own stores, competitor outlets, POIs and socio-economic characteristics of the consumers along with customer spotting data (see survey in table 1). In total, there were about 245 attributes available for 870 outlets at the high spatial aggregation level of municipalities and sub municipalities. The socio-demographic attributes, the competitor situation and number of POIs were mapped to 13 different trade areas for each outlet because the different location factors were supposed to take different effects on different spatial levels. The sales of selected outlets between December 2004 and November 2005 were used to train the data mining model and as comparison for the Huff-results.

Geometrical data	
Point geometries of own and competitor stores	
Trade area polygons	13 time and distance-depending trade areas for every outlet
Points of Interest	123 000 POIs of 3000 different branches
Street network geometries	
Units of the official area structure	
Postcode areas	About 2000 units
Attribute data	
Data about own stores	- sales floor - number of parking places - opening and closing dates - total sales between December 2004 and November 2005 - location type and quality
Data about competitor stores	- opening dates - shop types according to sales floor - type of business - estimated turnover according to federal state and competing enterprise
Market data	- Demographics (inhabitants according to age, sex and nationality) - Households (type and size) - Buying power (general and food specific) - Building activity and habitation (occupancy, monthly expenses, floor space, equipment) - Employment and education - (profession of sustainer and his occupational position, income branch, highest academic qualification, commuter structures) - Land allocation
Customer spotting data	
Home addresses of about 160 000 customers	

Table 1 ***SVR-input data***

5. THE APPLIED MODELS AND EVALUATION CRITERIA

The Huff-prediction is based on the straight-line distances between outlet location and the centroids of the consumer sub areas, sales floor in square meters as operationalisation of store attractiveness and food specific buying power in administrative sub areas. In a first step, the model was calibrated with empirical patronage probability of 160 000 customers. By averaging the lambdas of all sub areas belonging to one outlet, an outlet-specific distance exponent was computed. For new store locations, the average lambda of the five nearest neighbouring stores with customer spotting data was taken to calibrate.

The Huff-model which was implemented in this case study calculates sales according to the following formula:

$$U_j = \sum_{i=1}^{n} \frac{A_j * d_{ij}^{-\bar{\lambda}}}{\sum_{j=1}^{n} A_j * d_{ij}^{-\bar{\lambda}}} * Kk_i \quad (5)$$

The SVR-prediction was based on data about the following location factors:

- Car accessibility
- Sales area and location type
- Competition (number, predicted sales and absorption of buying power)
- Cannibalisation
- Shopping linkage behaviour
- Customer structure (Proportion of transient and regular customers)
- Customer-related socio-demographic attributes as described above
- Derived geographical attributes (population density, centrality, geographic coordinates, tourism potential)

Finally, 52 different explanatory attributes were chosen for the regression model. To meet the precondition of linearity the data was logarithmised. According to formula (4) the sales for a new outlet was calculated as follows:

$$exp\left(b + \sum_i w_i x_i' \right) \quad (6)$$

where b is the regression constant, $\sum_i w_i x_i'$ is the weighted sum of the logarithmised 52 attributes x', and exp being the exponential function with base 10.

In statistics there exist several measures to evaluate the performance of forecasting models. However, a forecasting model should at least produce better predictions compared to a naïve forecast and predicted values which correlate to the real value distribution. These are the minimum constraints for a prediction model. To verify if the models covered in this study meet these minimum constraints the following common error measures were adopted:

- RRSE (Root Relative Squared Error)
- Correlation coefficient for metrical data according to Bravais-Pearson
- ME (Mean Error)
- Standard deviation
- RMSE (Root Mean Squared Error)[1]

The RRSE can take values between 0 and infinity. Values smaller than one indicate an improvement in forecasting compared to a naïve predictor while values greater than 1 indicate a bigger error than the consulted naïve predictor. In this comparison, the arithmetic average has been used as naïve reference forecast. The ME quantifies the bias of a prediction. The bias is the difference between the average of the forecasts and the average of the observed distribution. It indicates whether an estimator tends to over- or underpredict systematically. The standard deviation is a measure of

[1] Formulas for RMSE, Standard deviation and ME see Walther, 2005.

precision. Measures of precision compare different estimators of the same population and can be defined as statistical variance of an estimator. Measures of accuracy display differences between the predicted and the observed values. In this study, the RMSE was used as accuracy measure.

6. DISCUSSION OF RESULTS

Table 2 shows the results of the Huff prediction (ME, standard deviation and RMSE in Euro).

Evaluation Criteria	**Value**
Number of outlets	864
Correlation	0.013
ME	-17 105.905
Standard deviation	247 970.828
RMSE	283 350.025
RRSE	2.03

Table 2 Prediction results for the Huff-model

In general, the Huff prediction does not satisfy the minimum constraints for a prediction model as stated in section 5: a correlation between real and predicted sales does not exist (see also the scatterplot in figure 4, real sales are depicted on x-axis, predicted sales on y-axis) and the RRSE indicates even a worse prediction accuracy compared to a naïve predictor. This means that even a naïve forecast produces better sales forecasts for potential new sites than the Huff-model for this use case. The low ME of 17.000 € means that real and predicted averages of the distribution do not differ very much, which indicates that the Huff-model is useful in predicting average sales while the high RMSE – which is sensitive to outlier prediction - shows that extremely good or extremely bad locations systematically tend to be under- respectively overpredicted. Additionally, the Huff-model slightly tends to underpredict sales systematically (as shown by the negative sign of the ME). Additionally, the model shows a bad precision because the standard deviation of the prediction is very high.

Table 3 and figure 5 show the results of the SVR-prediction (applied for a sample of 870 outlets). First of all, there is a significant correlation between real and predicted sales as well as a forecasting improvement compared to the naïve prediction (RRSE < 1). As in the case of the Huff-model the bias is relatively small but the sales are systematically underpredicted.

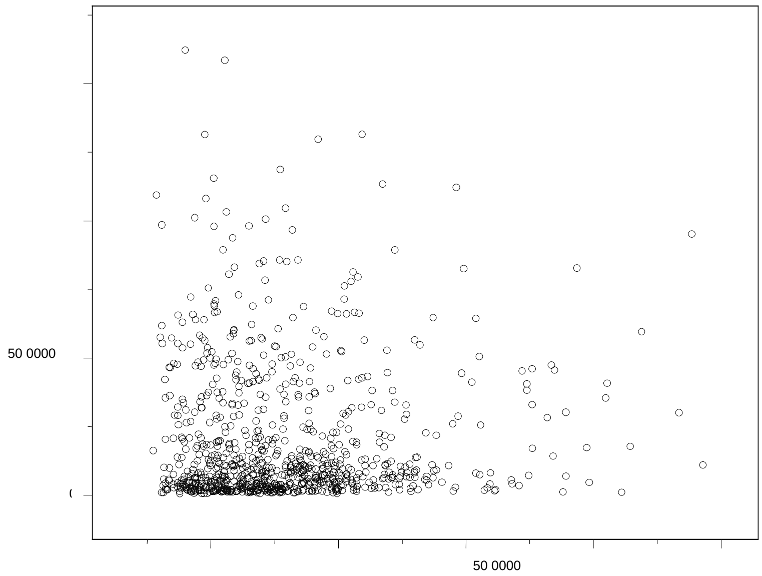

Figure 4 Scatterplot Huff-prediction

The RMSE is significantly lower compared to the Huff-prediction. This means that the SVR-model is better in predicting extraordinarily good or bad locations. And finally, the precision of the SVR-model is significantly better than that of the Huff-prediction.

Evaluation Criteria	**Value**
Number of outlets	870
Correlation	0.728
ME	-20 384.684
Standard deviation	95 815.158
RMSE	98 072.306
RRSE	0.702

Table 3 Prediction results of the SVR-model

There are two interpretations of these results, which are grounded in the modelling process itself and its exploitation of an empirical database. The modelling process of the Huff-model is strongly determined and involves a number of presuppositions. In contrast, the KDD-modelling process is iterative and more inductive. In the steps of task analysis and preprocessing, the data mining task is defined and suitable data is identified and prepared for data mining. During the phase of postprocessing the developed model is evaluated. In case the prediction results are not suitable, the analyst steps back into preprocessing and develops new features which are tested for significant impacts on the target value.

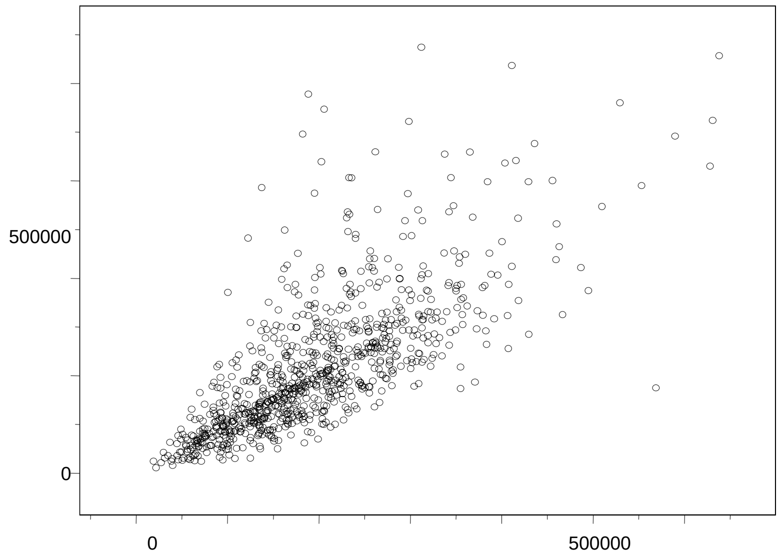

Figure 5 Scatterplot of SVR-prediction

This remodelling also took place in our use case. The first SVR-model, whose prediction results were not satisfying, did not take the described geographical attributes like geographical location or tourism into account. After an intensive analysis of sales data and possible influencing factors, a nonlinear dependency between sales and tourism was discovered. In the second phase of preprocessing, this feature was operationalised and given to the second data mining process. As table 4 shows, the geographical aspects in the form of tourism indicators have a significant impact on predicting sales of this food store chain.

Rank	**Operationalisation**
6	Number of Hotels and guesthouses
31	Ratio number Hotels and restaurants / population
52	Number of second- and holiday homes

Table 4 Operationalisation of the touristic feature and ranks

Therefore, an iterative modelling process is preferable to a linear one because it allows including explorative knowledge which was hidden in the data. The second reason why the data mining model is superior to the deductive Huff model is the extensive regress on empirical data. This means not only the use of a broad variety of socio-economic data and competitor characteristics, but also the exploitation of empirical sales data. Hence systematic errors in the explaining variables do not propagate to the predicted sales, because the SVR-model is trained with the empirical sales population.

7. CONCLUSION

The empirical comparison shows that potential sales for new outlets can be predicted very precisely with inductive methods of spatial data mining. The prediction accuracy of a SVR model is considerably better than the forecasts with the classical and widely used Huff-model. As could be shown, geographic knowledge which was discovered in the data mining process plays an important role in improving the forecast. Hence the improvements of the data mining approach justify the high input demand. The method is supposed to be applicable for chain stores like supermarkets or chemists trading convenience goods and hence operating in outlet-dominated markets. To conclude it must be stressed, that SVR may not be the best solution for retail sales forecasting in all cases. There may be other cases where kNN, Neuronal Nets or other techniques of DM or artificial intelligence may perform considerably better than SVR depending on the quality and quantity of input data or on the products or size of the company under investigation. Nevertheless, the authors expect that models developed by DM and artificial intelligence in general will always outperform the classical deductive models in terms of forecasting accuracy by the reasons shown in this example case.

BIBLIOGRAPHY

Applebaum, W., Methods for determining store trade areas, market penetration and potential sales. Journal of Marketing Research, 3/2005, 127 – 141, 1966.

Craig, C.S., Ghosh, A., McLafferty, S., Models of the retail location process. A review. Journal of Retailing, Vol 60, 5 – 36, 1984.

Fittkau, D., 2004 Beeinflussung regionaler Kaufkraftströme durch den Autobahnlückenschluß der A49 Kassel - Gießen. Zur empirischen Relevanz der "New Economic Geography" in wirtschaftsgeographischen Fragestellungen. Dissertation, Geographisches Institut Georg-August-Universität Göttingen, pp. 251. (German)

Gunn, S.R., Support Vector Machines for Classification and Regression. Technical Report: Faculty of Engineering, Science and Mathematics. School of Electronics and Computer Science. Southampton,1998.

Hernandez, T., Bennison, D., The art and science of retail location decisions. International Journal of Retail & Distribution, Vol. 28, No. 8, 357 – 367, 2000.

Huff, D.L., Defining and estimating a trading area. Journal of Marketing, Vol. 28, 34 – 38, 1964.

Kotschedoff, M., 1976 Sozialphysikalische Modelle in der regionalen Handelsforschung. Berlin-Verl., ISBN 3-87061-082-4, pp. 232. (German)

Megee, M., Forecasting economic base or structure by regression analysis. The Professional Geographer, Vol. 20, No. 1, 16 – 22, 1968.

Nakanishi, M., Cooper, L.G., Parameter Estimation for Multiplicative Interactive Choice Model: Least Squares Approach. Journal of Marketing Research, 11, 303 – 311, 1974.

Reilly, W.J., 1939 The Law of Retail Gravitation. - New York.

Rinzivillo, S. et al., Knowledge Discovery from Geographical Data. In: Gianotti, F., Pedreschi, D. (ed.). Mobility, Data Mining and Privacy. Springer: 243 – 263, 2008.

Vapnik, V., Golowich, V., Smola, A., Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. In Mozer, M., Jordan, M., Petsche, T. (ed.). Advances in Neural Information Processing Systems. Cambridge: 281 – 287, 1997.

Walther, B.J., The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography, Vol. 28**,** 815 – 829, 2005.