European Research Studies Journal
Volume XXIV, Special Issue 2, 2021

pp. 335-345

Methods of Analyzing Consumer Behavior Based on Multi-Source Data

Submitted 03/04/21, 1st revision 24/04/21, 2nd revision 25/05/21, accepted 15/06/21

Paweł Rymarczyk[1], Piotr Bednarczuk[2], Ryszard Nowak[3], Tomasz Cieplak[4]

Abstarct:

Purpose: *The aim of the article is to develop a system for analyzing processes and data from various data sources based on machine learning methods.*
Design/Methodology/Approach: *For data analysis, models for forecasting the level of sales and the pipeline gathering operations for the preparation of features, data, model training and its verification were designed. The data analysis pipeline is built of stages related to the preparation of features and data preparation. The learning process requires a specific division of data. The data set has been divided into three subsets, i.e., the training and validation data used in the learning process and the test subset used to verify the quality of the model.*
Findings: *The results of the conducted research show that the use of this type of analytical methods allows for the creation of new business processes, adaptation of services and goods to customer requirements, or the appropriate location of products on the retail space in order to optimize the time of shopping (especially taking into account the pandemic situation).*
Practical Implications: *The models presented in the article can be used by combining sales systems and behavioral data related to the movement of customers in the area of the sales space, where it is possible to build systems that allow optimization of orders, the way of arranging goods and other customer behavior patterns.*
Originality/Value: *A novelty is the construction of a multi-source model for data analysis, where appropriate predictive models were built to predict the level of sales with the use of machine learning algorithms.*

Keywords: *Machine learning, business data processing, data acquisition, prediction.*

JEL codes: *C51, C53, C80.*

Paper type: *Research article.*

[1] *Corresponding Aauthor, Netrix Group sp. z o.o. , Lublin, Poland, pawel.rymarczyk@netrix.com.pl*
[2] *University of Economics and Innovation in Lublin, Lublin, Poland, piotr.bednarczuk@wsei.lublin.pl*
[3] *University of Economics and Innovation in Lublin, Lublin, Poland, ryszard.nowak@wsei.lublin.pl*
[4] *Management Faculty, Lublin University of Technology, Poland, t.cieplak@pollub.pl*

1. Introduction

In today's complex business world, companies need to pioneer ways to differentiate themselves from other market participants by becoming more cooperative, effective, precise, and flexible. They must be able to quickly respond to the needs of the market and changes taking place in it (Paolanti *et al.,* 2018). Depending on the company's competitive advantage, which may be novelty, price, great website content or social media presence, specific online strategies must be applied to reach the desired market. Many companies have found that the data they store and the way they use it can build a market advantage. Data and information are becoming the basic resources for many organizations (Moreira, Ferreira, and Seruca, 2018; Sadowski, 2019).

For example, it is estimated that Walmart databases located in the cloud contain a total of over 40 PB and process 2.5 PB of data every hour, which includes information about customer behavior and preferences, network and device activity, and data on market trends (Garcia and Red, 2020). Business models are based on data from a variety of external sources, internal data warehouses and Internet resources. They form a process base that is a resource for computational intelligence algorithms. The functionality of intelligent decision support and business process modeling consists in managing strategic information, analyzing information from the direct and indirect business environment and influencing its strategic development.

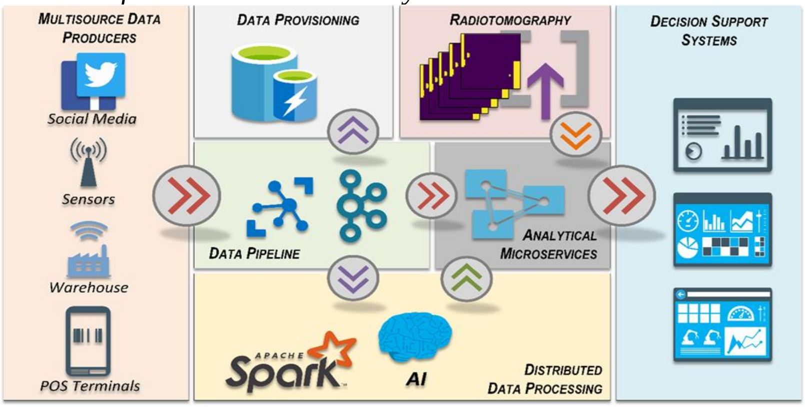
The traditional approach to system development focuses on building applications, not services. Systems built in this way are not flexible and difficult to scale (Newman, 2019). Nowadays, one of the main issues is how to capture the frequently changing needs and expectations of users and support those with dynamic business processes. In addition, the design of analytical systems must take into account the fact that there are many data channels that may need to be accessed and incorporated into further research. One of the key features of the cloud is on-demand access and secure access to scalable data mining resources. Cloud-based analytical systems provide unlimited access to various types and sources of data. This gives us the ability to process large amounts and a wide range of data with high speed and truthfulness, which means Big Data solutions.

The appearance of new technologies and data collection tools is now opening up areas of research on understanding customer-business relationships and interactions. Interesting from the point of view of the development of enterprises is the interdisciplinarity, which allows you to gain experience in many fields, at first glance not related to the management or organization of the enterprise - we are talking about physics or biology. Basic sciences currently offer many opportunities to study the processes taking place in enterprises, similar to those that take place, for example, in nature (Nowotny *et al.,* 2016). Unfortunately, there is currently no coherent approach to the use of multi-source models in the scientific literature and business practice, or they are used only selectively, which means that their full diagnostic and predictive potential is not used. In this context, currently relatively easily available customer data from sources such as social networks, digital media, mobile telephony, online games, online shopping, etc., empower business analysts to find a more precise, holistic analytical meta-model that would enable a detailed diagnosis of phenomena. driving modern business (Arsenault, 2017, Livingstone, 2019). As research results show, this process can be strongly enhanced thanks to artificial intelligence methods (Clark *et al.*, 2020).

2. Methodology and Results

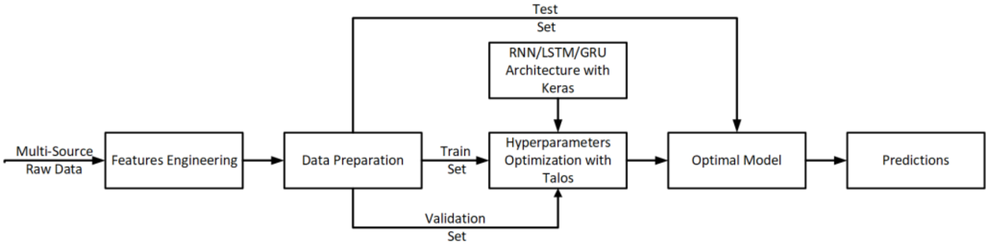
The development of the model for forecasting the level of sales in the future was carried out in an orderly manner in the form of a multi-source data analysis model (Figure 1) and a pipeline gathering operations for preparing features, data, model training and verification. The data analysis pipeline is built of stages related to the preparation of features and data preparation. The learning process requires a specific division of data. The data set has been divided into three subsets, i.e. the training and validation data used in the learning process and the test subset used to verify the quality of the model. The pipeline is shown in Figure 2, and a detailed description of the preparation of features and data is described in the following paragraphs.

Figure 1*. Model of multi-source data analysis*



Source: *Own creation.*

Figure 2*. Diagram of data analysis pipeline*



Source: *Own creation.*

This step is essential in the data analysis process. At this stage, when analyzing the formulation of the problem, various hypotheses are generated. Hypotheses are generated in such a way as to favor the expected result. The goal is to build a predictive model that allows you to predict the level of sales. The main idea at this stage is to identify product characteristics and product presentation venues that can affect sales. Since forecasting is based on products and product venues, we can break down these hypotheses into "Product Level Hypotheses" and "Store Level Hypotheses."

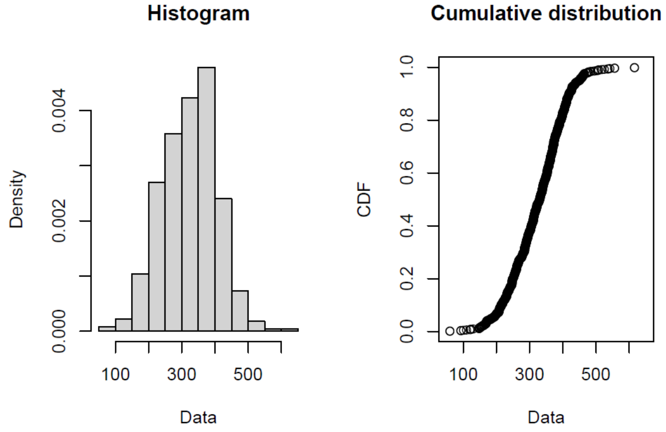
There are only a few hypotheses about the product level that can affect sales, they are, brand, packaging, exhibition space, promotional offers, visibility in the exhibition area, advertising in social networks, while the group of hypotheses about the way of presenting the product that may have sales impact can include, for example, city size, population, exhibition space, competitors, location, customer behavior, marketing, atmosphere, etc. For example, branded products have higher sales compared to other products, because customer trust in the brand is relatively higher, which leads to an increase in sales of this particular brand. Likewise, if the electronic presentation of products is well prepared and handled by reliable and courteous employees, they are expected to have higher sales. When a business problem is considered, it aims to achieve greater accuracy by changing and implementing different models. The dataset that has been collected covers furniture sales data from 2017 to 2019 and includes information on 63 products that are displayed in the physical exhibition space and on the company's website.

In order to properly understand the data and prepare them for the learning process, an analysis of the similarity of the empirical distribution to subsequent theoretical distributions was performed. If the distribution in the original data does not differ significantly from the theoretical distribution under consideration, it is possible to replace the data with the data in the given distribution with its appropriate parameters. This can significantly contribute to the improvement of the quality of the trained model and, above all, affect the very process of learning the neural network.

In the data set under consideration, the variables Web_View and Move_Coef were taken into account, which represent, respectively, the number of page views on the given day and the visit rate based on the RTI, which determines the customer traffic in the vicinity of subsequent collections in the store. The study presents the process of searching for the distribution that best matches the distribution of the Web_View variable. The Move_Coef variable remained unchanged in further analyzes due to the lack of a significant fit of its distribution to any of the basic theoretical distributions.

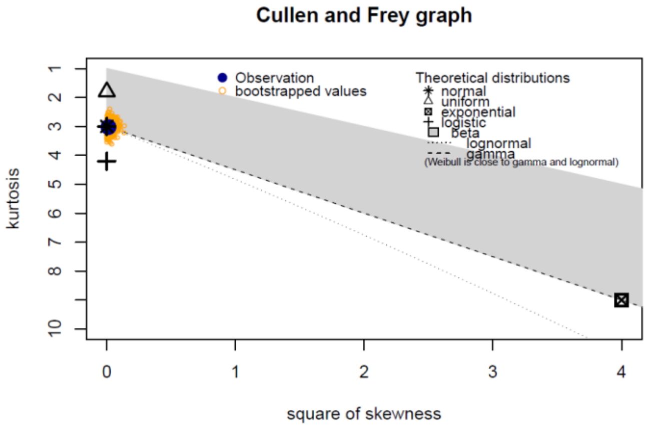
At the beginning, basic statistics and graphs showing the distribution of the feature were analyzed, as well as the Cullen-Frey diagram, which is often helpful in the task of identifying the appropriate theoretical distribution. The distribution of the Web_View variable is shown in Figure 3. Figure 4 shows the Cullen-Frey plot. On its basis, however, it cannot be determined what distribution the considered variable will probably have. Therefore, successive distributions should be examined in order to determine the most appropriate one.

Figure 3*. The empirical distribution of the Web_View variable*



Source: *Own creation.*

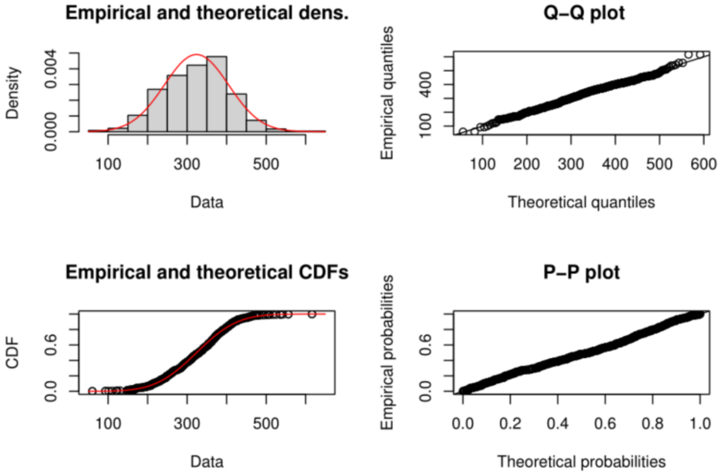
Figure 4*. Cullen-Frey plot for the Web_View variable*



Source: *Own creation.*

Thus, the smallest number of impressions in the tested time series was the number 60 and the largest 616. The relationship between the median and the mean and the value representing the skewness suggest some left-hand asymmetry. At this stage, we will not delve further into the data statistics, but will proceed to the search for the theoretical distribution by comparing successive distributions with the empirical distribution using the fitdistrplus library of the R package. The non-parametric Kolmogorov-Smirnov test will be used to test the compliance of the found distribution with the theoretical distribution. This test checks whether the distribution in the analyzed population of a random variable differs from the assumed theoretical distribution. Thus, the null hypothesis assumes that the distributions do not differ significantly. For the borderline significance level α, if the p-value resulting from the test is greater than α, then we confirm the null hypothesis. Otherwise, we reject the null hypothesis and say that the distribution of the variable differs significantly from the theoretical distribution under study. In our considerations, α = 0.05 was adopted. The first of the considered distributions is the normal distribution. Using the fitdist function from the fitdistrplus package, we obtain the parameters of the distribution most suited to the data. The graphs obtained are shown in the figure below (Figure 5).

Figure 5. *Fitting the Normal Distribution*

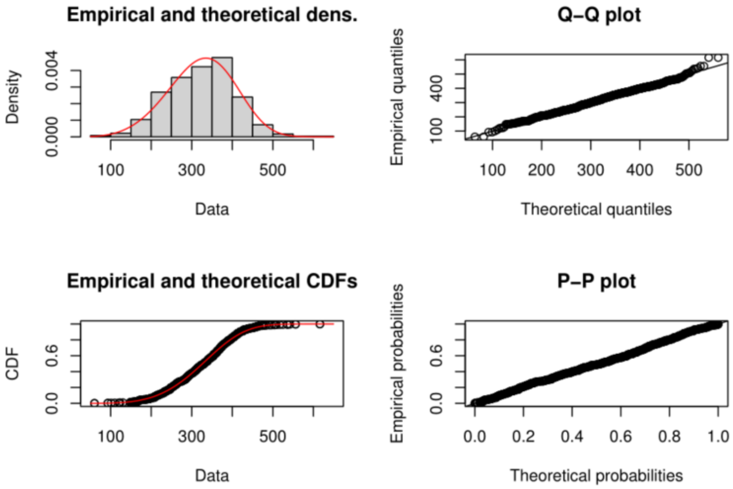


Source: *Own creation.*

Based on the above visualizations, it can be said that the normal distribution is noticeable here - the exceptions are outliers and the previously mentioned left-hand asymmetry. Therefore, it remained to check the results of the Kolmogorov-Smirnov test for the distribution determined in this way. The value of p-value = 0.03666 <0.05 obtained from the test makes us reject the null hypothesis, so the distribution of the Web_View variable differs significantly from the normal distribution. Similar procedures have been performed for many other distributions, but only the Weibull distribution will be presented. Proceeding as before, the Weibull distribution parameters were determined using the fitdist. This distribution seems to compensate to some extent for the impact of outliers. The K-S test gives the result p-value = 0.3857. This means that at the significance level α = 0.05 there are no grounds for rejecting the null hypothesis, so the distribution of the Web_View variable does not differ significantly from the Weibull distribution with the shape parameter 4.451642 and the scale parameter equal to 354.546389.

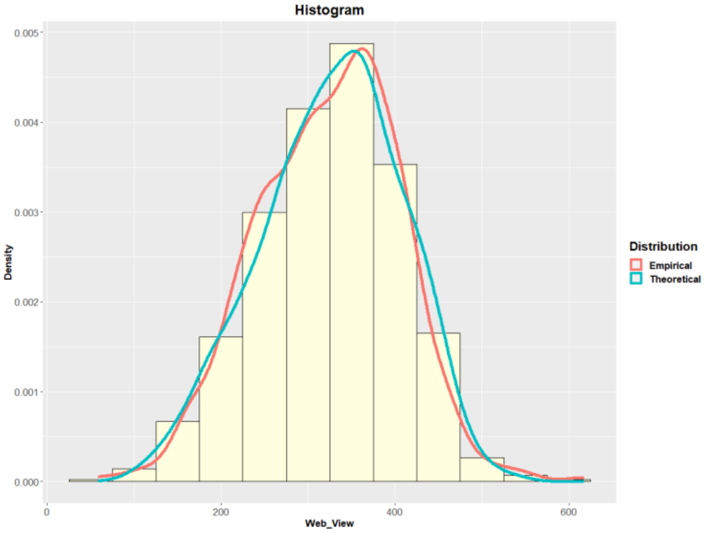
Based on the considerations and analyzes presented above, the Weibull distribution (Figure 6) turned out to be the best in terms of matching the distribution to the distribution of the Web_View variable. Figure 7 shows the histogram of the examined variable with the empirical distribution and theoretical distribution adjusted to it, the lack of differences being confirmed by the Kolmogorov-Smirnov test.

Figure 6. *Weibull distribution fit plots*



Source: *Own creation.*

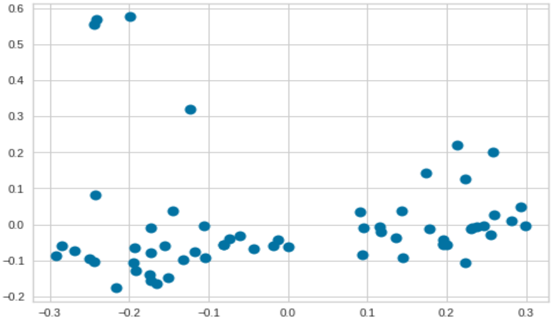
Figure 7. *Empirical and theoretical Weibull distribution*



Source: *Own creation.*

Due to the characteristics of the assortment sold by the examined company, the sales data must be divided into groups. The first division is made on the basis of the collection, i.e. a single project entering the market and treated as a whole (Figure 8). Note the length of the collection's existence on the market. For the purposes of creating predictive models, the collection must exist on the market for at least several years so that there is enough sales data to build a training, validation and test set on its basis. Then, in the selected collection, the assortment was divided into groups characterized by specific similarities. The Multidimensional Scaling method was used for this task.

Figure 8. *Distribution of elements of one collection*



Source: *Own creation.*

- *The Multidimensional Scaling method*

To present the similarities and differences of between product demand we employ the multidimensional scaling (MDS). It is a statistical technique, which allows us to visualize the similarities of groups of objects. Differences between analyzed groups are contained in a distance matrix $\left[d_{ij}\right]_{1\le i,j\le k}$. The elements of matrix are defined as follows:

$$d_{ij} = \left|cor\left(X_i, X_j\right) - 1\right| \qquad (1)$$

where X_i and X_j denote demands of these products. When the demands are strongly correlated and correlation ratio is positive $cor\left(X_i, X_j\right)$ tends to 1, relationship between demands is perfectly and $\mathrm{X_i}$ increases as $\mathrm{X_j}$ increases), then distance between demands of products tends to 0. In opposite when the demands are strongly correlated and correlation ratio is negative $cor\left(X_i, X_j\right)$ tends to -1, relationship between demands is perfectly and $\mathrm{X_i}$ increases as $\mathrm{X_j}$ decreases) then distance between demands of these products tends to 2.

The multidimensional scaling tends to locate the objects as points in space, where the similar elements are located close together. The multidimensional scaling seeks the points $z_i \in \mathbb{R}^2$, $1 \le i \le n$ that correspond to objects. By solution of the task:

$$\min_{z_1,\ldots,z_n} S(z_1, z_2, \ldots, z_n) \qquad (2)$$

we estimate the points corresponding to groups. Objective function:

$$S(z_1,z_2,\ldots,z_n) = \sum_{1\le i,j\le n} \left(D_{ij} - \|z_i - z_j\|\right)^2 \tag{3}$$

is called a stress function, $\|z_i - z_j\|$ is an Euclidean norm.

- *K-means cluster analysis*

The K-means clustering method consists in partitioning an unlabelled data set into non-overlapping clusters and determining the centers of the clusters. Initially, we establish the number of clusters K. Let $C = \{z_j\}_{1\le j\le n}$ denote a set of points corresponding to demands of products $z_j \in \mathbb{R}^2$. The task of the clustering method is to divide the entire C set into subsets called clusters $C_1, \ldots, C_K$, where $C_1 \cup \ldots \cup C_K = C$ and $C_i \cap C_j = \emptyset$ for $i \neq j$ and $1 \le i,j \le K$. For cluster C_k, $1 \le k \le K$, we within-cluster variation is defined as follows:

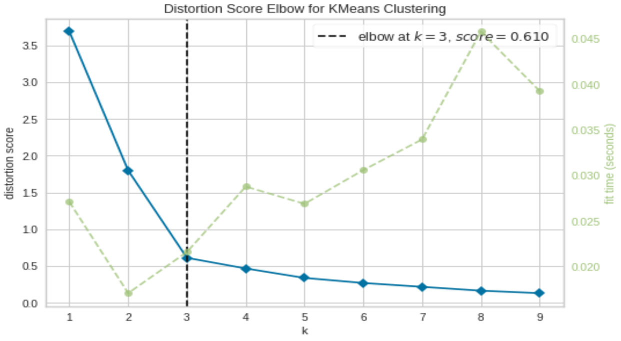
$$W(C_k) = \frac{1}{|C_k|} \sum_{i,j\in C_k} \|z_i - z_j\| \tag{4}$$

where $\|z_i - z_j\|$ is an Euclidean norm, $|C_k|$ - number of objects in k –th cluster. The main idea of clustering consists in dividing the data set into clusters so that the sum of within-cluster variations will be as small as possible. To determine the clusters we must solve the optimization task:

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} W(C_k) \tag{5}$$

The verification of the optimal number of groups was done using the elbow rule (Figure 9).

Figure 9*. Selection of the number of clusters using the elbow rule*



Source: *Own creation.*

3. Conclusions

The aim of the research work was to develop an innovative prototype of a system for analyzing consumer behavior. The authors of the system adopted the main goal of integrating and correlating data from multiple sources. In principle, the data may come from the subsystems of traffic and customer location monitoring as well as from devices used to transmit data from POS systems. The data obtained in this way will be analyzed (data mining) in order to find dependencies in the analyzed processes. The effects of using this type of analytical methods is the possibility of creating new business processes, adapting services and goods to customer requirements or appropriate location of products on the retail space in order to optimize the time of shopping (especially taking into account the pandemic situation).

Other main issues considered in the analysis are, identifying different groups of customers visiting stationary and online stores, understanding the specific needs and preferences of each segment, offering appropriate services to meet customer needs. The researchers' response to the needs of the retail industry was to develop effective tools for customer segmentation. There have been many studies using different types of data. Customer segmentation is the process of dividing heterogeneous customers into homogeneous groups based on common attributes and is necessary to more effectively serve different customers with rich sets of diverse customer preferences.

By combining data from sales systems and behavioral data related to the movement of customers in the area of the sales space, it is possible to build systems that allow you to optimize orders, how to arrange goods and other customer behavior patterns.

References:

Arsenault, A.H. 2017. The datafication of media: Big data and the media industries. International Journal of Media and Cultural Politics, 13(1-2), 7-24.

Clark, A., Zhuravleva, N.A., Siekelova, A., Michalikova, K.F. 2020. Industrial Artificial Intelligence, Business Process Optimization, and Big Data-driven Decision-Making Processes in Cyber-Physical System-based Smart Factories. Journal of Self-Governance and Management Economics, 8(2), 28-34.

Garcia, G.J.V., Red, E.R. 2020. Information Systems Utilization of an Organization: The Case of Walmart Inc. College of Computer and Information Science.

Livingstone, S. 2019. Audiences in an age of datafication: Critical questions for media research. Television & New Media, 20(2), 170-183.

Moreira, F., Ferreira, M.J., Seruca, I. 2018. Enterprise 4.0 – the emerging digital transformed enterprise? Procedia Computer Science, 138, 525-532. https://doi.org/10.1016/j.procs.2018.10.072.

Newman, S. 2019. Monolith to Microservices: Evolutionary Patterns to Transform Your Monolith. O'Reilly Media.

Nowotny, H. et al. 2016. Investigating interdisciplinary collaboration: theory and practice across disciplines. Rutgers University Press.

Paolanti, M., Liciotti, D., Pietrini, R., Mancini, A., Frontoni, E. 2018. Modelling and Forecasting Customer Navigation in Intelligent Retail Environments. Journal of Intelligent & Robotic Systems, 91(2), 165-180. https://doi.org/10.1007/s10846-017-0674-7.

Sadowski, J. 2019. When data is capital: Datafication, accumulation, and extraction. Big Data & Society, 6(1). https://doi.org/10.1177/2053951718820549.