Flash Crowd Management in Beyond 5G Systems

Valentin Rakovic(✉), Hristijan Gjoreski, Marija Poposka, Daniel Denkovski, and Liljana Gavrilovska

Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia
{valentin,hristijang,poposkam,danield,liljana}@feit.ukim.edu.mk

Abstract. Wireless network (radio) virtualization and its synergy with ML/AI-based technologies is a novel concept that can efficiently address problems of legacy networks, such as flash crowds. This paper discusses the integration aspects of intelligence-based technologies with Sate-of-the-Art end-to-end reconfigurable, flexible and scalable network architecture, capable of handling demands in flash crowd scenarios. The presented results, demonstrate that advanced solutions based on ML can significantly improve the network proactivity and adaptivity by reliably predicting flash crowd scenarios. The results also show that in case of low dataset fidelity, conventional statistical models are a more suitable option.

Keywords: Flash crowd · Radio virtualization · Machine learning

1 Introduction

Conventional network architectures are characterized by static deployment and configuration, rendering them uncapable for managing geographical and spatio-temporal variations of users' capacity demand. The network inflexibility presents a challenging problem in the attempt to satisfy the service demands for the increased number of flash crowd scenarios in the recent years, especially emergency situations such as terrorist attacks and natural disasters. The overall scenarios' outcome depends on the ability to reliably and promptly exchange information between the first team responders and the victims.

The strict requirements imposed by the flash crowd scenarios demand a sheer transformation from future mobile systems. These scenarios introduce specific service requirements that traditional network architectures cannot provide on-the-fly and in a manner that operators would be willing to support (i.e. low-cost, efficient and real-time service demand satisfaction). However, the emerging 5G systems [1] and the associated aspect of radio virtualization promises to address these challenges.

Some of the focal aspects of 5G are the radio network virtualization and softwarization that enable flexible, scalable, agile network architectures, which address the demands

The original version of this chapter was revised: The family name of the first author was corrected. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-78459-1_31

D. Perakovic and L. Knapcikova (Eds.): FABULOUS 2021, LNICST 382, pp. 48–57, 2021.
https://doi.org/10.1007/978-3-030-78459-1_4

in flash crowds. Radio virtualization [2] allows isolated coexistence of multiple virtual radio networks on the same physical infrastructure, and has ability to dynamically create heterogeneous virtual networks. Network softwarization, represented through Software Defined Networking (SDN) [3, 4], enables network programmability by decoupling data and control planes. It also decouples the network functions from the hardware where they commonly run and implements them in software manner. It brings an opportunity to place the network functions in a virtualized environment (i.e. cloud platform).

Recently the radio virtualization has started to evolve towards the concept of Open-RAN (O-RAN) [5]. O-RAN is fostering more open and smarter radio access networks by relying on openness and Intelligence. To address the complexity issues, operators and vendors cannot rely on conventional human intensive means of deploying, optimizing and operating the mobile networks. Instead, the mobile networks must be able to facilitate new intelligence-based technologies (i.e. ML/AI), hence facilitating automated operational network functions that will reduce the operational costs.

This paper discusses the aspects of flash crowd scenarios and presents a promising solution for its effective mitigation. Specifically, the paper presents the requirements for an end-to-end reconfigurable, flexible and scalable network architecture, capable of addressing the demands in flash crowd scenarios. The emphasis is on the design and performance analysis of a self-autonomous network entity called Virtual Resource Manager (VRM) capable of orchestrating the different underlying radio access technologies, in case of flash crowd occurrence.

The paper is organized as follows. Section 2 provides an insight on the flash crowd specific, and consequently the underlying system requirements and design. Section 3 focuses on different algorithms for flash crowd prediction, based on the underlying network information. Section 4 presents the performance analysis of the algorithms presented in Sect. 3. It also discusses the potential applicability and deployment for real-world scenarios. Section 5 concludes the paper.

2 Flash Crowd Scenarios

2.1 Flash Crowd Aspects

Flash crowd scenarios are becoming important due to the increased requirement for providing a reliable communication and ubiquitous access infrastructure for a large number of active users and devices. The given scenarios are characterized by a substantial increase of user connectivity and/or traffic demand, as a result of high concentration of users per unit area (e.g. sport events, concerts etc.) or by a high volume of required user connections (e.g. tele-voting, emergency situations, etc.). The flash crowd scenarios can affect many system's characteristics, such as cell densification and self-organization, resource reconfiguration, system outages and failures, etc. The most demanding and important flash crowd scenarios involve emergency situations that arise either from out of natural (e.g. earthquakes, floods) or man-made (e.g. terrorist attacks, industrial accidents, transportation failures) disasters.

The fifth generation of mobile systems, 5G, envisions a paradigm shift by introducing very high carrier frequencies with massive bandwidths, dynamic softwarization and virtualization, high reliability, low latency and very high number of connected users. A

number of ongoing activities focus on system architecture and initiates standardization of various aspects of 5G, such as ML/AI integration into the system [6]. However, applying advanced algorithms such as ML/AI in flash crowd-aware future wireless networks is very limited. The remainder of the section presents the generic system architecture capable of addressing the issues related to flash crowds in mobile systems.

2.2 System Design

In order to provide reliable and efficient communication in flash crowd scenarios the wireless systems, such as 5G, should focus on virtualization and autonomous decision-making algorithms (e.g. based on ML/AI). Virtualization and autonomous decision-making can facilitate optimized virtual resource allocation for wireless network environments, leveraging a highly flexible and adaptable wireless network.

The flash crowd-aware system should aim at supporting the operation of various underlying wireless technologies and to dynamically and optimally allocate the cloud hardware resources onto the available wireless network resources, depending on the environment and the traffic demands. There have been several recent works that propose the integration of virtualization and autonomous decision-making algorithms in mobile systems. However, only the FALCON-based wireless system focuses on the aspects of flash crowd scenarios [7]. The remainder of the section will specifically elaborate on the FALCON system architecture and its flash crowd-aware design.

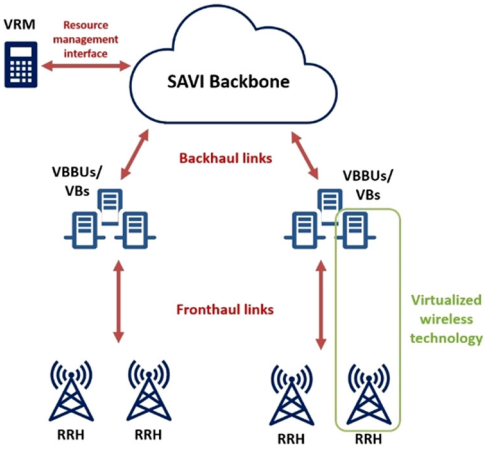

Fig. 1. Generic FALCON architecture

Figure 1 depicts the generic FALCON system architecture. The generic FALCON architecture integrates the Radio Access Network and the backbone core network in a single reconfigurable system entity. The RAN segment incorporates the Cloud-RAN (C-RAN) paradigm, comprising of Remote Radio Heads (RRHs) that are located at remote sites. They transmit and receive wireless signals, but do not perform any base-band signal processing. The digital baseband signal is forwarded, via a high speed/low

latency fronthaul links, to the Virtual BaseBand Units (VBBUs)/Virtual Base Stations (VBSs) housed in the centralized processing pools. The VBBUs/VBSs control the RRHs and perform the baseband processing of the digital signals to retrieve the essential data from the respective virtualized wireless communication system. The communication between the RRHs and VBBUs can be provided by Software Defined Network (SDN) equipment that can dynamically adapt to the underlying scenario requirements and conditions. Moreover, the FALCON architecture incorporates aspects of Multi-Access Edge Computing (MEC) [8] to support a distributed cloud deployment (C-RAN). The MEC aspect improves the performance, the scalability and the agility of the C-RAN platform by bringing the centralized processing pools closer to the wireless nodes [9].

Based on the data traffic/user connectivity demands, as well as on the virtualized wireless network technology, the Virtual Resource Manager (VRM) computes the hardware resources in the MEC processing pool (CPU load, memory allocation, etc.) providing optimal performance and resource allocation. The FALCON architecture integrates the proposed MEC based C-RAN with the SDN-based SAVI Backbone platform via backhaul links. SAVI acts as the core facilitator of the backbone communication and services in the C-RAN system, leveraging advanced network management [10].

The VRM is the focal entity in the FALCON system architecture responsible for the optimal system reconfiguration under particular flash crowd occurrences. The VRM's main objective is to proactively detect and predict the underlying flash crowds in the network. The following section presents the key VRM functionalities responsible for the flash crowd prediction.

3 Flash Crowd Prediction

In order to facilitate the detection process, the VRM utilizes particular prediction algorithms that can be applied over a large plethora of scenarios. Specifically, these scenarios vary in the type and content of available network information (i.e. dataset) that the VRM utilizes for its decision-making process.

Specifically, the scenarios and the applied algorithms vary with respect to the information size in the dataset. When utilizing large amount of historical information from the dataset, the optimal approach is to exploit advanced prediction algorithms based on ML. However, in many real-world situations mobile operators operate with limited historical data, rendering the ML algorithms unusable. In this case it is more efficient to apply conventional statistical algorithms.

The remainder of the section elaborates on the dataset aspects and the optimal algorithms that can be applied in the VRM for the particular scenarios of interest.

3.1 The Dataset

The traffic characteristics over time are an important aspect of cellular networks in consideration of resource provisioning, traffic engineering and system optimization. There has been some progress in revealing temporal dynamics and spatial inhomogeneity of cellular traffic, however there is quite a limited knowledge about traffic dependence, and statistics analysis about the load of the network: number of users, number of data

packets, size of the data transfer, etc. One of challenges is the lack of relevant data collected in real-life conditions. The City Cellular Traffic Map (C2TM) dataset [11, 12] is the only one that contains such data, and additionally provides analysis on week-long traffic generated by a large population of people in a median-size city of China.

The dataset is collected over 8 continuous days [11, 12]. The data represents request-response records extracted from HTTP traffic at the city scale, consisting of individuals' activities, with accurate timestamp and location information indicated by connected cellular base stations (BS). The hashed International Mobile Subscriber Identity (IMSI) detects each individual. To conserve the traffic characteristics and also to preserve user privacy, only hourly statistics at base-station granularity is available. This means a maximum of N $\times$ M records (N the number of base stations, M the number of recording hours).

The dataset contains two types of data: traffic and topology. The former provides hourly traffic statistics for each base station, while the latter stores the relative topology of underlying cellular network. The relative location of base station is in longitude/latitude form to facilitate some analysis with standard geographic processing about great circle distance.

In our analysis we used only the traffic related data, since we were interested only in the traffic frequency characteristics. This data contains:

- timestamp
- base station ID
- number of active users associated with specific base station and hour
- number of transferred packets associated with specific base station and hour
- number of transferred bytes associated with specific base station and hour.

In total, there are 13.269 base stations and 1.625.680 data rows (hours), in the presented dataset. However, the dataset is not evenly distributed, i.e., there is missing data for some of the base stations. In the first step, we filtered the dataset in in order to leave only the base stations that contain complete data, i.e., 192 h of data (8 days). This resulted in keeping 1.983 base stations, i.e., 380.736 data rows (hours).

Then, we used this dataset to develop two approaches that predict the number of users one hour in advance. The first one is based on machine learning regression algorithms and predicts the number of users for an individual base station, i.e., it uses the characteristics of that particular base station to predict the number of users in the next hour. The size of the dataset is 380.736 data samples. The second approach uses other more generic statistical algorithms to predict the number of users on aggregated level, i.e., in all of the base stations. It utilizes only 192 data samples (the same as the number of hours), which is too small dataset that can be applied to advanced ML.

3.2 Individual Base Station Prediction

This approach is based on machine learning, regression models that predict the number of users at a particular base station in one-hour interval. The model takes as input the current and the historical (last 24 h) traffic data from the particular base, then calculate numerous statistical features, and finally provide the prediction for the next hour.

The first step in the development of the machine learning model is the features extraction. In our case we use as input the current and the historical (last 24 h) traffic data from the particular base station, and then calculate numerous statistical features. We calculated the following statistical features for the last 24 h for each of the characteristics: number of users, number of packets, and number of bytes: mean value, max value, min value, kurtosis, standard deviation, skewness. Additionally, as a separate feature we use the previous value of each of the 3 characteristics.

In, the next step we trained 4 regression models: Decision Tree [13], KNN [14], Random Forest [15], and Xtreme Gradient Boost [16] models. Additionally, we used Dummy Regressor [17] as a baseline model. This model is simple statistical model that predicts a constant, i.e., the mean value of the users in the train data.

3.3 Aggregated Prediction

Since cellular traffic manifests usual time series behavioral model, including seasonal and trend patterns, statistical models can be utilized for forecasting the future outcomes in regard to past statistics. In order to identify a pattern or formula that may apply in future data prediction, many researchers have developed statistical analysis methods for time series [17]. The paper analyses 4 statistical analysis models, i.e. Moving Average (MA) [18], Exponential Smoothing (ES), i.e., Single Exponential Smoothing (SES) and Double Exponential Smoothing (DES) [19] and SARIMA [20]. These models are trained to predict the number of users on aggregated level for all base stations.

Moving Average Model

MA is a simple model that aims to predict trend-cycle elements by smoothing past data. A future data point is estimated as average of the k previous equally weighted data points. The MA data point estimate Y_t^{MA} is given by the following formula:

$$Y_t^{MA} = \frac{X_t + X_{t-1} + \ldots + X_{t-(k-1)}}{k} \tag{1}$$

where X_t is data point value at time t and$k$ represents the number of past data points used in the MA model. Choosing this number properly is of great importance. Specifically, very small values will not capture the trend present in the time series, while very large values will impair the forecasting accuracy.

Exponential Smoothing Model

Since the most recent data observations yield the best reference for the future, there is need for schemes that have decreasing weights in time, as the observations get older. ES models are example for weighted schemes that have exponentially decreasing weights. Although there are many ES models, they all weight recent observations more significantly than the older values by using smoothing parameters. In our case, we use SES where the most recent observation is the most weighted DES, which takes into account the possibility of trend and seasonal components by introducing one more term in the observation. The equation for next predicted value Y_{t+1}^{SES} for SES model is given by:

$$Y_{t+1}^{SES} = \alpha X_t + (1 - \alpha) Y_t^{SES}, \tag{2}$$

where Y_t^{SES} is present predicted value, X_t is present actual value and α is smoothing parameter,$0 \leq \alpha \leq 1$.

In case of DES, the forecast value Y_{t+1}^{DES} is given by:

$$Y_{t+1}^{DES} = \alpha X_t + (1-\alpha)(Y_t^{DES} + b_t, \tag{3}$$

$$b_{t+1} = \beta\left(Y_{t+1}^{DES} - Y_t^{DES}\right) + (1-\beta)b_t, \tag{4}$$

where Y_t^{DES} is present predicted value, X_t is present actual value and α is level smoothing parameter ($0 \leq \alpha \leq 1$), β is trend smoothing parameter ($0 \leq \beta \leq 1$) and b_t s estimate of the slope of the time series at time t.

SARIMA Model

Seasonal Autoregressive Integrated Moving Average, known as SARIMA is actually modified ARIMA that supports data with seasonal component. The model consists of trend elements and seasonal elements that can be chosen through thorough analysis of autocorrelation function and partial autocorrelation functions of the time series. Trend elements are represented by the following parameters:

p: Trend autoregression order.
d: Trend difference order.
q: Trend moving average order
On the other hand, there are 4 seasonal parameters:
P: Seasonal autoregressive order.
D Seasonal difference order.
Q: Seasonal moving average order
m: seasonal length in the data

This hyperparameters are used to specify the seasonality (S), autoregression (AR), differencing (I) and moving average (MA), so *SARIMA (p, d, q)(P, D, Q)m* is applied to the time series x_t with the following equation [21]:

$$\Phi\left(L^m\right)\phi(L)\Delta^d\Delta_m^D x_t = \theta_0\Theta\left(L^m\right)\theta(L)w_t, \tag{5}$$

where L is the lag operator and w_t is assumed to be Gaussian white-noise process with mean zero and variance σ^2. Δ^d and Δ_m^D are assumed to be difference operator and seasonal difference operator. This operators aim to transform the non-stationary time series x_t to the stationary process x_t^*:

$$x_t^* = (1-L)^d(1-L)^D y_t. \tag{6}$$

Further, $\phi(L)$ and $\theta(L)$ are defined as the following polynomials in the lag operator:

$$\phi(L) = 1 - \phi_1 L - \ldots - \phi_p L^p, \tag{7}$$

$$\theta(L) = 1 + \theta_1 L + \ldots + \theta_q L^q. \tag{8}$$

The seasonal polynomials $\Phi(L^m)$ and $\Theta(L^m)$ in the lag operator are defined as:

$$\Phi\left(L^m\right) = 1 + \Phi_1 L^m - \ldots - \Phi_p L^{Pm}, \tag{9}$$

$$\Theta\left(L^m\right) = 1 + \Theta_1 L^m + \ldots + \Theta_p L^{Qm}. \tag{10}$$

The following section analyses the prediction performance of the presented algorithms, in the case of flash crowd occurrences.

4 Performance Analysis

This section evaluates the prediction accuracy and reliability of the presented algorithms, for both the individual and aggregate base station predictions. The analysis is performed by examining the Mean Absolute Error (MAE), as a metric for evaluation and comparison between the different algorithms. To evaluate the models the dataset was divided 70% for training, and 30% for evaluation.

Figure 2 depicts the MAE results for the ML-based algorithms used in the per base station prediction. The figure shows that all of the models significantly outperform the baseline. This suggests that ML algorithms are more reliable and suitable for flash crowds' predictions compared to conventional prediction algorithms, when the system has large volumes of historical information. As a result, the ML algorithms were able to learn a model that uses the characteristics of the dataset to better predict the number of users. The figure also shows that the best performing model is the Random Forest. It achieves MAE of 7.1 users, which means that on average the model will under- or over-estimate the number of users by 7.

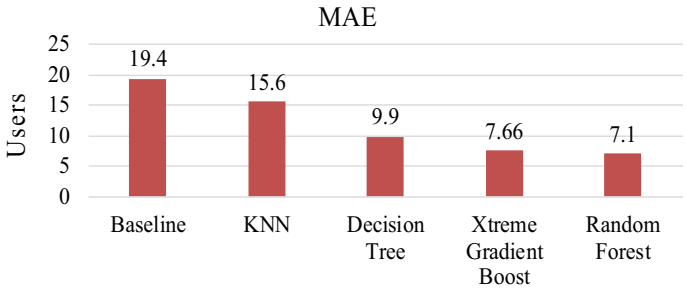

Fig. 2. Individual prediction of the number of users. Comparison of models.

Random Forest and Xtreme Gradient Boost induce the highest computational complexity of all analyzed models. In particular, they consist of hundreds of Decision Trees. In scenarios where the virtual RAN is computational capabilities are limited or computationally heavy loaded, Decision Tree can be more suitable because of their significantly lower computational complexity and relatively high accuracy. Another advantage of the Decision Tree is that it is understandable model, i.e., an expert can check and evaluate the model visually. This is something that we plan to further exploit in future work.

Figure 3 depicts the results and the comparison between the statistical models used for the aggregate base station prediction. The results show that the MA model shows worst performance, since it is the simplest time series model. Between the SES and DES, it is clear that DES model performs better due to the fact that it takes into account the trend and seasonality characteristics of the time series. It is also evident that the SARIMA model outperforms all of the previous models as a result of its complexity and the numerous trend and seasonal parameters that play a big role in the prediction process.

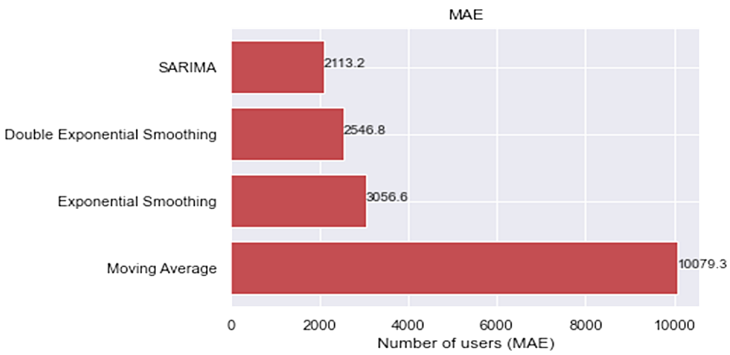

Fig. 3. Aggregated prediction of the number of users. Comparison of models

Taking into consideration that the average value of the data set in the aggregate prediction case, which is 56 815.4 users, SARIMA achieves less than 4% error in the prediction process. In the case of the individual base station prediction, the average number of users is 29, thus Random Forest achieves 24% error. However, the distribution of the number of users among the base stations varies a lot, i.e., the standard deviation is 28, resulting in the higher error percentile. For example, the Baseline model, that reflects the statistical predictors, has significantly higher prediction error compared to the Random Forest.

5 Conclusion

As a result of their inflexibility and static configuration, existing network architectures are failing to provide service in the most critical periods related to flash crowd situations. The 5G, as a promising technology, still do not address properly the utilization of the network virtualization and its dynamicity with respect to flash crowds.

This paper discusses the potential applicability of RAN virtualization and its synergy with new intelligence-based technologies, such as ML, that can address the issues related to flash crowd scenarios. The results in the paper clearly show that advanced solutions based on ML can significantly improve the network performance and its proactive adaptivity by reliably predicting flash crowd scenarios. The ML-based solutions are especially efficient when the available historical information is large and when the number of active users has significant variation. The paper also discusses that in case of low

dataset fidelity or limited computational capabilities, conventional statistical models are a more suitable option.

Future work will extend the analysis to specific flash crowd scenarios, such as emergency situations. Moreover, it will focus on practical implementation of the proposed algorithms and foster real-world trials.

References

1. 5GPPP Architecture Working Group: View on 5G Architecture (2017)
2. Chowdhury, N.M.M.K., Boutaba, R.: A survey of network virtualization. Comput. Netw. Int. J. Comput. Telecommun. Netw. **54**(5), 862–876 (2010)
3. Kreutz, D., Ramos, F.M.V., Veríssimo, P.E., Rothenberg, C.E., Azodolmolky, S., Uhlig, S.: Software-defined networking: a comprehensive survey. Proc. IEEE **103**(1), 14–76 (2015). https://doi.org/10.1109/JPROC.2014.2371999
4. European Telecommunications Standards Institute: Network Functions Virtualisation (NFV), NFV#17 Plenary meeting, Bilbao, Spain (2017)
5. Gavrilovska, L., Rakovic, V., Denkovski, D.: From cloud RAN to open RAN . Wireless Pers. Commun. **113**, 1523–1539 (2020). https://doi.org/10.1007/s11277-020-07231-3
6. ITU Focus Group on Machine Learning for Future Networks including 5G (2020). https://www.itu.int/en/ITU-T/focusgroups/ml5g/Pages/default.aspx
7. Marinova, S., et al.: End-to-end network slicing for flash crowds. IEEE Commun. Mag. **58**(4), 31–37 (2020). https://doi.org/10.1109/MCOM.001.1900642
8. Gavrilovska, L., Rakovic, V., Denkovski, D.: Aspects of resource scaling in 5G-MEC: technologies and opportunities. In: IEEE Globecom 2018 (2018)
9. European Telecommunications Standards Institute: MEC in 5G networks, Sophia Antipolis, France (2018)
10. Kang, J.M., Lin, T., Bannazadeh, H., Leon-Garcia, A.: Software-Defined Infrastructure and the SAVI Testbed. In: Leung, V., Chen, M., Wan, J., Zhang, Y. (eds.) Testbeds and Research Infrastructure: Development of Networks and Communities. Springer, Cham (2014)
11. The dataset: https://github.com/caesar0301/city-cellular-traffic-map
12. Chen, X., Jin, Y., Qiang, S., Hu, W., Jiang, K.: Analyzing and modeling spatio-temporal dependence of cellular traffic at city scale. In: 2015 IEEE International Conference on Communications (ICC) (2015)
13. Hssina, B., Merbouha, A., Ezzikouri, H., Erritali, M.: A comparative study of decision tree ID3 and C4.5. Int. J. Adv. Comput. Sci. Appl. **4**(2), 13–19 (2014)
14. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R., Verleysen, M.: K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing **72**(7–9), 1483–1493 (2009)
15. Breiman, L.: Random forests. UC Berkeley TR567 (1999)
16. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y.: Xgboost: extreme gradient boosting. R package version 0.4-2, pp. 1–4 (2015)
17. Makridakis, S., Wheelwright, S.C., Hyndman, R.J.: Forecasting Methods and Applications. Wiley Student Edition, 3rd edn. ISBN 9780-471-532330
18. Box, G.E.P., Hunter, J.S., Hunter, W.G.: Statistics for Experimenters: Design, Discovery, and Innovation, 2nd edn. Wiley, Hoboken (2005). 13 978-047 J -71813
19. Gardner, E.: Exponential smoothing: the state of the art—part II. Int. J. Forecast. **22**, 637–666 (2006). https://doi.org/10.1016/j.ijforecast.2006.03.005
20. Fuller, W.A.: Introduction to Statistical Time Series, 2nd edn. Wiley, New York (1996)
21. Brockwell, P.J., Davis, R.A.: Time Series: Theory and Methods. Springer, New York (1991)