UNIVERSITÀ
DI TRENTO

Department of Information Engineering and Computer Science

Bachelor's Degree in Informatica

Final Dissertation

Title

Subtitle (optional)

Supervisor
.....

Student
.....

Academic year 2019/2020

Contents

Summary **2**

1 State of the art **2**
1.1 Literature review . . . 2
1.2 Time series forecasting . . . 3
1.2.1 Exponential smoothing . . . 3
1.2.2 Artificial Neural Networks . . . 3
1.2.3 ARIMA models . . . 4
1.2.4 TBATS models . . . 4
1.3 Queueing theory . . . 4

2 Data analysis **5**
2.1 Inflow rate, dwell time and service time distributions . . . 5
2.2 Time series analysis . . . 5
2.2.1 Time series decomposition . . . 6
2.2.2 Inflow rate autocorrelation . . . 6

3 Methodology **6**
3.1 Model overview . . . 6
3.2 Inflow rate forecast . . . 6
3.2.1 Persistence model . . . 6
3.2.2 Naïve model . . . 6
3.2.3 Artificial Neural Network model . . . 7
3.2.4 ARIMA model . . . 7
3.3 Arrival rate forecast . . . 7
3.4 Queue length forecast . . . 7
3.4.1 Service rate approximation . . . 8
3.4.2 Outflow rate approximation . . . 8
3.5 Checkouts optimization . . . 8
3.5.1 Opening/closing fluctuations control . . . 8

4 Results **8**

5 Conclusions **8**

References **8**

Summary

Sommario è un breve riassunto del lavoro svolto dove si descrive l'obiettivo, l'oggetto della tesi, le metodologie e le tecniche usate, i dati elaborati e la spiegazione delle conclusioni alle quali siete arrivati.

Il sommario dell'elaborato consiste al massimo di 3 pagine e deve contenere le seguenti informazioni:

- contesto e motivazioni
- breve riassunto del problema affrontato
- tecniche utilizzate e/o sviluppate
- risultati raggiunti, sottolineando il contributo personale del laureando/a

1 State of the art

This chapter discuss the available related literature. First, different approaches from similar researches are described. All this researches fall into the broader discipline of *Operational Research* (OR) and mainly into two specific sub-fields: *Time Series Forecasting* and *Queueing Theory*. The state of the art techniques for this fields are discussed in the following two chapters.

Terminology used:

- *Inflow/outflow rate*: number of customers that enter/exit the supermarket in a given interval.
- *Arrival rate*: number of customers that arrive at the checkouts in a given interval.
- *Service rate*: maximum number of customers that can be served by each checkout in a given interval.
- *Dwell time*: time spent by a customer in the supermarket.

1.1 Literature review

There are various studies that try to analyze and predict the queue length in different settings, but little was found about staff optimization.

Berman et al. (2004) [1] developed a system that manages the switching of workers between a "front room", where the checkouts are located, and a "back room", that is the shop. This system process real-time data about the count of customers either in front and back rooms, with the goal of minimizing the customer's waiting time. To accomplish this performance objective, they first defined a *waiting time threshold* and required it to not be exceed. Second, they defined a minimum *time-average worker complement threshold* to complete all the back room work. They used a $M/M/\infty$ queue model to determine the performance of the system with a given number of front room workers. With this model, they were able to suggest in real-time the optimal strategy of switching workers from one room to another, minimizing the average waiting time while satisfying the thresholds constraints.

Aksu H. (2018) [5] main goal was to optimize the idle time of the staff operating at the checkout while maintaining the waiting queue length under a predetermined threshold. He proposed a model based on the inflow of customers at the entrance, the number of customers entering the checkout area, the current waiting queue length and number of available checkouts. He divided the supermarket area

into a shop and a checkout area and used a video-based system to count the customers moving in each area. He then developed a prediction model to calculate a realistic forecast of customer dwell times. The inward flow into the checkout area was predicted using the inward flow into the shop, since the inward flow into the shop appear with a delay at the checkout area, and this delay is the dwell time. To consider non-standard events, the model was adjusted with real-time measured data, to increase the prediction accuracy. Queueing theory was then used to predict the average waiting time at the checkouts. Checkouts can be opened or closed based on a minimum and maximum acceptable queue length values. To avoid closing and opening too frequently, the system was able to decide in how many cases the defined waiting queue length can be exceeded, e.g. it is acceptable to exceed the target value for a short period of time.

1.2 Time series forecasting

A *time series* is a collection of observations made sequentially through time. *Time series forecasting* is the use of a model to predict future values based on previous observations.

A time series can be decomposed into three main components:

- *Trend* T_t: Linear/nonlinear, increasing/decreasing behavior of the series over time.
- *Seasonality* S_t: Repeating patterns or cycles behavior over time.
- *Residual* R_t: Variability of the observations not explainable by the model.

There are two types of decomposition: *additive*, where the time series would be written as $y_t = T_t + S_t + R_t$, and *multiplicative*, where $y_t = T_t \cdot S_t \cdot R_t$. The additive decomposition is the most appropriate if the magnitude of the seasonal fluctuations, or the variation around the trend-cycle, does not vary with the level of the time series. When the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series, then a multiplicative decomposition is more appropriate

The following sections describes different techniques that can be used to obtain a forecast of future values.

1.2.1 Exponential smoothing

In the *Simple exponential smoothing* technique, the forecasted values are based on a weighted average of the previous values, where the most recent observations are given more importance using larger weights [6].

A more complex implementation is the *Holt-Winters exponential smoothing* approach, which is able to decompose the time series into a level, trend and seasonal component, giving a more precise forecast. However, this method is unable to model more complex series with multiple seasonality patterns, as in a supermarket inflow rate.

Taylor's *double seasonal exponential smoothing* method [9] was developed to forecast time series with two seasonal cycles: a short one that repeats itself many times within a longer one.

1.2.2 Artificial Neural Networks

Artificial Neural Networks (ANN) allows complex nonlinear relationships between the target variable and its predictors. With time series data, lagged values of the time series can be used as inputs, in what is called an *Autoregressive Neural Network (AR-NN)*. With seasonal data, it is also useful to also add the last observed values from the same season as inputs [6]. In general, the input can be expressed as:

$$(y_{t-1}, y_{t-2}, ..., y_{t-p}, y_{t-m}, y_{t-2m}, ..., y_{t-Pm})$$

We use the notation AR-NN$(p, P, k)_m$ to indicate a neural network with p previous lagged values and P lagged values from previous seasonal cycles in the input, k neurons in the hidden layer, and a seasonal cycle of m intervals.

For forecasting two steps ahead, the result of one-step forecast can be used in the input.

1.2.3 ARIMA models

ARIMA (AutoRegressive Integrated Moving Average) models provide another approach to time series forecasting. Exponential smoothing and ARIMA models are the two most widely used approaches to time series forecasting, and provide complementary approaches to the problem. While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data.[6]

AutoRegressive means the model uses the relationship between an observations and some number of lagged previous observations to generate a linear regression model. *Integrated* refers to the use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary. A stationary time series has mean, variance and autocorrelation constant over time. *Moving Average* means that the model uses the dependency between an observation and the residual error from a moving average model applied to lagged observations.

The parameters of the ARIMA model are defined as follows:

- p: Autoregressive order, the number of lagged observations included in the model.
- d: Degree of differencing, the number of times that the raw observations are differenced.
- q: Moving average order, the size of the moving average window.

A simple version of the ARIMA model can be used to forecast only non-seasonal time series, but it can be extended to model seasonal patterns with the Seasonal ARIMA (SARIMA). There are four seasonal parameters that are not part of ARIMA that must be configured:

- P: Seasonal autoregressive order.
- D: Seasonal degree of differencing.
- Q: Seasonal moving average order.
- m: The number of time steps for a single seasonal period.

A SARIMA$(p, 0, 0)(P, 0, 0)_m$ model is equivalent to an AR-NN$(p, P, 0)_m$ model but with a restrictions on the parameters that ensure stationarity. However, only a single seasonal effect can be modelled with SARIMA. To model more seasonalities the SARIMAX model can be used, since it supports exogenous variables. The seasonal effects can be captured by Fourier terms and used as exogenous variables. This gives a better approximation and allows other exogenous variables to be considered (e.g. weather, holidays, ...) to further improve the accuracy.

1.2.4 TBATS models

An alternative approach proposed by De Livera et al. [2] uses a combination of Fourier terms with an exponential smoothing model and a Box-Cox transformation, in a completely automated manner. The main advantage of a TBATS model over a SARIMAX model with Fourier terms is that the seasonality is allowed to change slowly over time, while harmonic regression terms force the seasonal patterns to repeat periodically without changing. However, one drawback of TBATS models, is that they are slow to estimate, since they will consider various alternatives and fit different models.

1.3 Queueing theory

Queueing theory is the mathematical study of waiting lines, or queues. A queueing model is constructed so that queue lengths and waiting time can be predicted.

There are various types of queueing models, but the best representation of a supermarket queue is probably given by the $M/M/c$ model, that represent a system with an infinite queue capacity, where the inflow rate follow a Poisson distribution and the service times of the c servers follow an exponential distribution. The customers are served in FCFS order (First Come First Served).

The *utilization factor* $\rho = \frac{\lambda}{c\mu}$ describes the proportion of total service capacity being used in the system. If $\rho >= 1$, i.e. the arrival rate exceeds the service capacity, then the queue will grow

indefinitely, but if $\rho < 1$ the system has a stationary distribution and the average queue length can be calculated.

The main issue with this stationary approach in the context of supermarkets is that the arrival rate and service rate cannot be expressed with probability distributions with constant mean, since they are strongly time-dependent, thus non-stationary. Moreover, it could be possible that the utilization factor exceeds 1 for a short period of time (*overloading*), in which the queue length increases, and after that the system returns slowly to its equilibrium, when $\rho < 1$.

A common approach for dealing time-varying rates is the *Stationary Independent Period by Period* (SIPP) approach, where the analysis is conducted on small intervals [8]. For each interval, a different $M/M/c$ model with constant arrival rates and number of servers is created and solved with the stationary approach. However, Green et al. (2001) [4] shown that the commonly used SIPP approach is inaccurate for parameter values corresponding to many real situations, even when the time-dependent variations are small and especially for systems which operate near the critical load. Moreover, this stationary analysis requires the arrival rate to be strictly smaller than the service rate, i.e. $\rho < 1$ must hold for every interval, while many service systems in reality can be temporarily overloaded.

Green et al. (1991) [3] proposed an easy-to-compute approximation for determining long run average performance measures for multi-server queues with periodic arrival rates, the *Pointwise Stationary Approximation* (PSA). This approximation was obtained by computing the expectation of the performance measure over the time interval using the stationary formula with the arrival rate that corresponds to each point in time. They showed that this approximation is an upper bound for the expected number of customers in the queue. However the PSA can only be applied to systems where $\rho < 1$ hold, i.e. temporal overloading in some intervals is strictly forbidden [8].

Rider [7] proposed an approximation of the average queue length for a M/M/1 model with time-varying parameters. He found that the average queue length $Q(t)$ satisfies the equation:

$$Q(t) = \lambda(t) - \mu(t)[1 - P_0(t)]$$

where $P_0(t)$ is the probability of the server being in idle. This equation can be interpreted to mean that the queue will increase with the arrival rate and decrease with the service rate multiplied by an "efficiency factor" $1 - P_0(t)$, i.e. the probability that the server is busy at time t. $P_0(t)$ is not a priori known, thus an approximation must be used.

Stolletz (2008) [8] proposed an improvement of the SIPP approach, the *Stationary Backlog-Carryover* (SBC) approach. This approach was designed for systems with temporal overloading, thus the models of consecutive periods are no longer independent from each other. A *backlog* of work is measured in each period and propagated into future periods. Contrary to the SIPP approach that applies the $M/M/c/\infty$ model to each interval, SBC utilizes the $M/M/c/c$ model with an artificial arrival rate $\hat{\lambda}(t)$ that consists of both the original arrival rate $\lambda(t)$ and the backlog rate $b(t)$ of the previous period.

2 Data analysis

This chapter describes the analysis conducted. Two different dataset were analyzed...

2.1 Inflow rate, dwell time and service time distributions

The distribution of inflow rates, dwell time and service time have been investigated.

2.2 Time series analysis

Different time series analysis techniques were used to extract the main components of the time series.

2.2.1 Time series decomposition

Decomposing the time series into different components is useful to find additional patterns and trends and to validate the previous observations. The main goal was to identify trend variations among different seasons.

2.2.2 Inflow rate autocorrelation

To identify the number of previous values that directly influence the current value, the *autocorrelation* of the inflow rate has been calculated. This measurement is useful to identify a first set of relevant features to be used in the forecast model. Autocorrelation refers to the degree of correlation between the values of the same variables across different past observations in the data. It measure the linear relationship between a variable's current value and its past values. The autocorrelation values can range from -1 to +1, where +1 represent a perfect positive correlation (an increase in one of the value leads to a proportional increase in the other value), -1 represent a perfect negative correlation and 0 represent no linear correlation. The *partial autocorrelation* is the autocorrelation between two values after removing any linear dependence on the values between them.

Considering more lagged values, it was clear that the time series present two main seasonalities:

- a *daily* seasonality, with a cycle length of $\sim$ 150 lags ($\sim$ 24 hours), meaning that the periods with more or less traffic are nearly the same for each day.
- a *weekly* seasonality, with a cycle length of $\sim$ 1000 lags ($\sim$ 7 days), meaning that there is a strong relation between the inflow rates for the same days in different weeks.

3 Methodology

This chapter describes the methods used to build the prediction model.

3.1 Model overview

The final model is composed by different parts that cooperates to obtain the final result.

First, the measured inflow rate for the recent past is combined with a forecast of the inflow rate for the immediate future. This values, combined with a dwell time prediction, are used to obtain a prediction of the arrival rate at the checkouts. The arrival rate is then used with queueing theory methods to calculate the expected queue length for a given interval.

3.2 Inflow rate forecast

The inflow rate can be expressed as a time series, hence time series forecasting methods were used. As described in Chapter 1.2, different forecasting methods were tested and the best was selected based on the performance obtained.

3.2.1 Persistence model

The first model tested was used to define a baseline in performance. This gives an idea of how well the other models could perform on the time series, and if a model's performance are worse than this baseline it should no be considered. The baseline was obtained by a *persistence model*, were the last measured value y_{t-1} is used as forecast:

$$\hat{y}_t = y_{t-1}$$

3.2.2 Naïve model

As seen in Chapter 2, the arrival rate time series presents a strong seasonal component, so the forecast should be based on this repeating pattern. The forecast value was calculated as the average of the

values in the previous weeks for the same day and time:

$$\hat{y}_t = \hat{f}(t) = \frac{1}{N}\sum_{i=1}^{N} y_{t-iW}$$

where W is the number of intervals in a week and N is the number of previous week considered.

The model performance were worse than the baseline. Looking at the results, it is clear that this model is not able to consider the variations in the inflow rate that can be caused by external factors, e.g. holidays. To take in account this variations, a local drift was calculated using the forecasted errors value of the immediate past:

$$\hat{y}_t = \hat{f}(t) + \frac{1}{M}\sum_{i=1}^{M} \hat{y}_{t-i} - \hat{f}(t-i)$$

were M is the number of previous step considered.

3.2.3 Artificial Neural Network model

As described in Chapter 1.2.2, Artificial Neural Networks are a popular and effective time series forecasting technique. We can denote an autoregressive neural network as AR-NN$(p, P, k)_m$. Various configuration parameters were tested and the model with the best performance was used.

3.2.4 ARIMA model

3.3 Arrival rate forecast

Once an inflow rate forecasting model was defined, the measured inflow rate was combined with the predicted values and used to get a prediction of the arrival rate at the checkouts, because the inflow rate into the shop appears with a delay at the checkout area [5]. This delay is the customer's dwell time. For such a prediction, a probability density function $p_t(\tau)$ is required, where τ is the dwell time of a customer. In Chapter 2.1 the general distribution of the dwell time was presented and with the analysis of Chapter 2.2 it is clear that the average dwell time values are strongly dependent from time, thus a forecasting model for the dwell time was created. In Chapter 2.1 an approximation of the dwell time distribution was given by the Erlang distribution. This approximation was used to define the function $p_t(\tau)$ using an average of the measured data from the latest weeks as the distribution parameter and creating different values for different time intervals and days. Let $\hat{\lambda}(t)$ be the predicted arrival rate for the interval t, $p_t(\tau_i)$ the probability of having a dwell time $\tau_i = i \cdot \Delta t$, Δt the size of the intervals considered and y_{t-i} the inflow rate measured at time $t - i$. We can write:

$$\hat{\lambda}(t) = \sum_{i=1}^{K} y_{t-i} \cdot p_t(\tau_i)$$

where K is such that $p_t(\tau_K) = o(0)$

For a multi-step forecast for the next N intervals, we can write a linear system of equations:

$$\begin{bmatrix} \hat{\lambda}(t) \\ \hat{\lambda}(t+1) \\ \vdots \\ \hat{\lambda}(t+N) \end{bmatrix} = \begin{bmatrix} y_{t-1} & y_{t-2} & \cdots & y_{t-(N+1)} \\ y_t & y_{t-1} & \cdots & y_{t-N} \\ \hat{y}_{t+1} & y_t & \cdots & y_{t-(N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{y}_{t+(N-1)} & \hat{y}_{t+(N-2)} & \cdots & y_t \end{bmatrix} \begin{bmatrix} p_t(\tau_1) \\ p_t(\tau_2) \\ \vdots \\ p_t(\tau_N) \end{bmatrix}$$

3.4 Queue length forecast

This chapter describes the implemented model for forecasting the queue length at the checkouts. As described in Chapter 1.3, with the M/M/c model the required variables are:

- $\lambda(t)$: Arrival rate at the checkouts.

- $\mu(t)$: Service rate of the checkouts.
- $c(t)$: Total count of available (open) checkouts.

3.4.1 Service rate approximation

3.4.2 Outflow rate approximation

3.5 Checkouts optimization

3.5.1 Opening/closing fluctuations control

4 Results

5 Conclusions

Bibliography

[1] Larson R. Berman O. A queueing control model for retail services having back room operations and cross-trained workers. *Computers and Operations Research*, 31:201–222, 02 2004.

[2] Snyder R. D. De Livera A. M., Hyndman R. J. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011.

[3] Linda Green and Peter Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37:84–97, 01 1991.

[4] Soares J. Green L. V., Kolesar P. J. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49:549–564, 08 2001.

[5] Aksu H. Dwell time forecast and checkout optimisation in supermarkets. -, 2018.

[6] Athanasopoulos G. Hyndman R. J. *Forecasting: Principles and Practice*. OTexts, 2nd edition, 2018.

[7] Rider K. L. A simple approximation to the average queue size in the time-dependent M/M/1 queue. *J. ACM*, 23(2):361–367, 04 1976.

[8] Stolletz R. Approximation of the non-stationary m(t)/m(t)/c(t)-queue using stationary queueing models. *European Journal of Operational Research*, 190:478–493, 10 2008.

[9] James W. Taylor. Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, 19(4):715–725, 2003.