Vrije Universiteit Amsterdam, Faculteit der Exacte Wetenschappen
Centrum voor Wiskunde en Informatica Amsterdam

INTERNSHIP REPORT

Melania Călinescu

Forecasting and Capacity Planning for Ambulance Services

Supervisors: Prof. dr. Rob van der Mei (CWI)
Dr. René Bekker (VU)

Amsterdam
2009

Preface

The master programme Business Mathematics and Informatics at VU University Amsterdam is a multidisciplinary study aimed at improving business processes by applying a combination of methods based on mathematics, computer science and business management. The practical component that concludes the master study is a six-month internship.

I undertook my internship at Centrum voor Wiskunde en Informatica Amsterdam (CWI) in the Probability and Stochastic Networks Group. The assignment is part of a bigger project done in cooperation with ambulance service providers in the Netherlands and it regards the development of better forecasting techniques for the received call volumes and better capacity planning methods.

The experience gained through this research project will be very valuable in my future career. I would like to thank prof. dr. Rob van der Mei (CWI) and dr. René Bekker (VU) for giving me the opportunity of an internship at CWI. I am grateful for the creative ideas and useful remarks they provided during my work. Their challenging questions and guidance helped me grow both as a student and as a researcher.

Furthermore, I want to thank my colleagues at CWI for a great working atmosphere.

Last but not least, I would like to thank Jakub Pečánka M.Sc. for the support on the programming part of the project and for his comments on my thesis.

Amsterdam, August 2009.

Abstract

Today, only little is known about how to efficiently plan ambulance services. Key issues, such as uncertainty in demand (in this case, emergency call volumes) and supply (in this case, available vehicles and ambulance personnel), have to be addressed more thoroughly. The current paper provides insight in dealing with randomness in the ambulance service planning. The focus is on high-priority calls.

We analyze whether a Poisson approximation of the call arrival process is suitable and what distribution would reasonably fit the occupancy time data (the total time an ambulance is busy). Based on these new findings, we develop a simple capacity planning method. Through a statistical analysis we discover that the call arrival process displays a monthly seasonal pattern and a day-of-the-week pattern. We also find that an inhomogeneous Poisson process approximates reasonably well the call arrival process during the day. Consequently, a forecasting model for the call arrival process is developed. A similar analysis is conducted to identify patterns in the occupancy time. As a results, two distribution approximations are given: Erlang (5) and Hypo-exponential distribution, with the latter one being the more intuitive choice.

The Erlang-B model with time-dependent arrival rate and time-dependent mean occupancy time is used to design the capacity planning. Validation of the results is also provided and we conclude that the Erlang-B performs well.

Directions for further research include a time-dependent approach, which applies the occupancy time distribution approximations, the extension of the current results to incorporate other types of calls and the design of crews schedule according to the new capacity planning method.

The confidential sections of the current internship report are available upon request.

Contents

1 Introduction

1.1 About CWI

CWI (Centrum voor Wiskune en Informatica) is the national research institute for mathematics and computer science in the Netherlands. From its foundation in 1946, CWI has performed fundamental scientific research within these fields and maintained an active link with the society through the channel of knowledge: new ideas have been discovered, developed and transferred to various domains, such as telecommunication, stock market, public transport, internet, meteorology etc. As one of the founding members of the European Research Consortium for Informatics and Mathematics, CWI has set the standards for international cooperation between researchers, attracting talents from several countries world-wide.

The current strategy of the institute focuses more on practice, in the sense that researchers at CWI thrive to actively answer questions raised within the four most relevant society fields: Earth and life sciences (modeling, simulation and data analysis that accompany conventional experimentation for geological and biological research), Data explosion (models, methods and techniques necessary to manage, study and exploit the increasing amounts of data), Societal logistics (principles and methods that address crucial issues in the society: efficient and flexible organization of traffic and transport, commerce and public services) and Software as service (definition and deployment of standards and methods for the discovery, evaluation, combination and integration of services, without access to the underlying source code, within the context of the tremendous Internet growth)[1].

1.2 Motivation for the Current Research

Part of the research done within the Societal logistics topic concerns an efficient planning of ambulance services. Delivering high-quality service at affordable costs is of crucial importance, not only for ambulance service providers everywhere, but also for everyone of us that experience life-threatening situations. Today, only little is known about how to efficiently plan ambulance services. Handling the large costs generated by the acquirement and maintenance of the emergency medical service equipment and the assurance at the same time of highly qualified staff is a complex optimization problem that also has practical limitations such as data availability and computational resources. Therefore, there is need for the development of smart planning methods so that ambulance service providers could assure a high service quality level.

The optimization of various aspects of emergency medical service vehicle systems has been a very active area of research for applied mathematics and operations research. There has been a great deal of articles dealing with the development of models that provide support in the decision making process of key issues such as the scheduling of crews, bases locations, capacity and staffing of ambulance bases etc (see the following section for most relevant

[1]More information on the CWI research fields can be found electronically at *http://www.cwi.nl*

references in the field). Nevertheless, one aspect seems to lack from all these papers and that is the analysis of the stochastic nature of the input data. Neglecting the impact of uncertainty and assuming an a priori known deterministic demand (in this case, emergency call volumes) and supply (availability of vehicles and ambulance personnel) inevitably leads to inefficient planning of ambulance services.

The current paper provides insight in dealing with randomness in the ambulance service planning and, subsequently, develops appropriate quantitative models that help implement a more efficient planning of ambulance services.

The first step we make is to analyze whether a Poisson approximation of the call arrival process is suitable. Several questions arise from this direction, such as the existence of a daily effect, the occurrence of a moment-of-the-day effect, the presence of trend and/or seasonal patterns etc. and what the impact of these effects is on the rate at which calls arrive at the call center. As a short explanation, the existence of a daily effect means that the call volumes in different days of the week might be generated by Poisson distributions with different parameters. A moment-of-day-effect is detected when calls in the morning arrive at a different rate than calls in the afternoon, which suggests that an inhomogeneous Poisson process would be more adequate to model the call arrival process. Moreover, we check whether the accuracy of daily and hourly call volumes predictions is indeed improved through the addition of the special-effects parameters to the forecasting model.

The second step regards the travel time distribution (further denoted as the service time) from the base to any emergency site, which if known leads to the possibility of computing any steady state performance measure, as suggested in Restrepo (2008). Of course, the degree of complexity will increase (significantly for more elaborate models). However, simple prediction models could easily integrate this specification, especially since closed-form expressions are already known for a queueing model with a series of distributions of the service time. The mathematical analysis is accompanied by statistical tests (performed in `R`), carried out on the available data.

We proceed in the third and final step with the capacity planning by adapting models from the literature to our new findings (call arrival behavior and service time distribution) and carrying out the tailored forecasting procedures.

The most important result of the current research project is the simplicity of the developed models. As the performed statistical analysis has proved, the stochastic nature of the input data can be effectively captured through simple forecasting models. The predictions are obtained with high accuracy and in short computation times. Thus, our models build a sustainable source of forecasting techniques for further phases of estimating call volumes simultaneously over time and space, static and dynamic (re)deployment of ambulances among a number of bases.

1.3 Literature Review on Ambulance Service Planning

Ambulance service planning has stirred the interest of a great number of researchers. The literature on the topic is extensive, with developments in several directions, which include scheduling of crews, ambulance stations locations, and capacity and staffing of ambulance stations for both static and dynamic deployment of ambulances. The static ambulance deployment problem refers to the optimization problem of allocating a fixed number of ambulances among a set of bases, with the ultimate goal of ensuring the best possible medical

outcomes for patients. The dynamic ambulance deployment problem refers to the real-time relocation of idle ambulances among a set of stations. Through this strategy, repositioned idle ambulances can compensate for those that are busy, hence unavailable to respond to incoming calls. For a review of the research done within the static and dynamic ambulance deployment topics, we refer the reader to Restrepo (2008), which presents both previous methods and also new approaches to finding solutions for the two optimization problems.

Due to the nature of the available data, this paper focuses on the staffing of one single ambulance base. It addresses, however, the aspect of randomness present in the call arrivals and travel times. Although improving the ambulance service planning models in the context of stochasticity is a rather new approach, there are several papers that are relevant to our work. For instance, the analysis of special-day effects and seasonal patterns within the call arrival process, was tackled by Channouf et al. (2001) in their attempt of developing forecasting techniques using time-series models.

Another key result is presented in the case study of Erkut et al. (2009) on data coming from the Emergency Medical Services system of Edmonton, Canada. The paper evaluates the performance of several maximum coverage optimization models in terms of uncertainty in response times and ambulance availability. The maximum coverage models deal with the optimal number of ambulance stations so that the average response times to any demand node is within a preset limit.

Ingolfsson et al. (2008) also incorporates randomness for both travel times and ambulance availability but adds a random delay component, which accounts for the activity prior to travel to the scene. The model minimizes the number of ambulances needed to provide a specified service level for a set of (existing or planned) ambulances stations with known locations. A similar approach was developed in Erkut et al. (2008) by incorporating a survival function into existing covering models. A survival function is a monotonically decreasing function of the response time of an emergency medical service vehicle to a patient that returns the probability of survival for the patient.

More on how travel times depend on distances and how to use this dependence to improve coverage can be found in Budge et al. (2008). Daytime patterns for travel times are also discussed and included in the model.

The more complex problem of crews scheduling in the context of static ambulance deployment with random travel times was treated in Erdoğan et al. (2009). They develop a search algorithm that solves the static ambulances location problem so that the expected coverage is maximized, while considering probabilistic response times.

Another complex planning tool for the static deployment of ambulances is presented in the case study of Henderson and Mason (2005), developed for St. John Ambulance Service provider in Auckland, New Zealand. The novelty of their tool comes from the direct use of real data as recorded in a database (trace-driven simulation), the use of a detailed time-varying travel model for modeling travel times in the simulation, and the development of a geographic information system which provides a spatial visualization of the data.

Our method attempts to combine results from these papers into a simple but effective staffing rule. It is based on the incorporation of the randomness present in call arrivals and travel times, but in addition it provides a thorough statistical analysis of the data behavior. This case study confirms the Poisson distribution is an adequate approximation of the call arrival process (which is the most usual assumption in theory). It also reveals a new approximation for the travel times random duration (through a hypo-exponential distribution). To obtain the capacity planning, the well-known Erlang-B formula is applied. A validation

procedure analyzes the impact of the proposed planning in practice.

1.4 Structure of the Report

The remainder of the report is organized as follows. Chapter 2 provides some background knowledge concerning the main mathematical aspects that are included in our analysis. In Chapter 3, we develop the forecasting model for the daily and hourly call volumes, using insights provided by statistical analysis on historical data. Chapter 4 provides the data analysis and the approximation of the travel times distribution and Chapter 5 the proposed staffing schedule on an hourly basis. The performance of the models is evaluated through several error measurements. Conclusions and directions for further research are presented in Chapter 6.

2 Background Knowledge

The main mathematical aspects covered in this paper are summarized in the following sections in order to facilitate a better understanding of the results of the current research.

2.1 Poisson Process

The counting process $N(t), t \geq 0$, is said to be a *Poisson process having rate* λ, $\lambda \geq 0$, if

- $N(0) = 0$.
- The process has independent increments.
- The number of events in any interval of length t is Poisson distributed with mean λt. That is, for all $s, t \geq 0$

$$\mathbb{P}(N(s, s+t) = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \text{ for } k = 0, 1, 2 \dots .$$

Since most of the events modeled through Poisson processes are arrivals to a system (telephone call requests at a switchboard, page view requests to a website etc.), we further refer to our events as arrivals.

An equivalent characterization of the Poisson process is by the interarrival times, which are then independent and exponentially distributed with parameter λ. The equivalence can be seen by looking at the probability that there are no events within $[s, t]$:

$$\mathbb{P}(\text{next event after } s+t | \text{event at } s) = \mathbb{P}(N(s, s+t) = 0) = e^{-\lambda t}.$$

Thus, the time until the kth arrival has as distribution a sum of exponentially distributed random variables, which is commonly known as a Gamma or Erlang distribution with shape parameter k and scale parameter λ.

The above theoretical considerations concern a homogeneous Poisson process. In this case, given a number of arrivals $N[0, T]$, then they are uniformly distributed in $[0, T]$. For the inhomogeneous case we abandon this assumption and instead we assume that the arrival time is determined according to a distribution with a piece-wise continuous density f on $[0, T]$.

Define $\gamma = \mathbb{E}N(T)$, the expected number of events in $[0, T]$ and $\lambda(t) = f(t)\gamma$, the *the rate function*. Then the number of arrivals for an interval $[s, t]$, with $t \leq T$, $N(s, t)$ has a Poisson distribution with parameter $\gamma \int_s^t f(u)du$ and $\mathbb{E}N(s, t) = \int_s^t \lambda(u)du$. Again, arrivals in disjunct intervals are independent. Note that if $\lambda(t)$ is constant, then we have a homogeneous Poisson process.

For the interarrival times, the time until the next arrival after a fixed point in time is characterized by the rate function $\lambda(t)$. Take X_1 as the time until the first arrival, then

$$\mathbb{P}(X_1 > t) = \mathbb{P}(N(t) = 0) = e^{-\int_0^t \lambda(s)ds}.$$

Thus X_1 can have any distribution, depending on the rate function $\lambda(t)$.

For alternative definitions and additional properties of the Poisson process, we refer the reader to Ross (1997) and Koole (2009).

2.2 Erlang Loss Model

Several traffic models exist which share their name with the Erlang unit of traffic. They are formulae which can be used to estimate the number of servers required in queueing systems.

One of the most commonly used traffic models is the *Erlang B model*. It determines how many servers are required in a system if the traffic intensity (in Erlangs) is known. The model assumes that all blocked calls are immediately cleared from the system and it is based on the *Erlang formula*:

$$B = \frac{\frac{\rho^N}{N!}}{\sum_{i=0}^{N} \frac{\rho^i}{i!}},$$

where N denotes the capacity of the system and ρ the traffic intensity, which is defined as the product of the call arrival rate and the mean service time. The output of the Erlang formula is the percentage of blocked calls. An important property of the Erlang formula is its insensitivity to the distribution of the service time. This makes the Erlang formula both simple to apply and robust to changes in the traffic characteristics.

Another well-known model is the *Erlang C model*, which assumes that all blocked calls stay in the system until they can be handled. This model is most often applied to the design of call centers agents scheduling where, if calls cannot be immediately answered, they enter a queue. An extension of the *Erlang C model* is the *Erlang A* model, which includes abandonments, meaning that the calls waiting in the queue may leave the queue without being served.

2.3 Hypo-exponential Distribution

The *hypo-exponential* distribution or the generalized Erlang distribution is a continuous distribution with applications in queueing theory and more generally in the field of stochastic processes. The Erlang distribution is a series of k exponential distributions, each with rate μ. The hypo-exponential is a series of k exponential distributions each with their own rate μ_k. If we have k independently distributed exponential random variables X_k, then the random variable $X = \sum_{i=1}^{k} X_i$ is hypo-exponentially distributed.

Its name derives from the fact that it has a coefficient of variation smaller than one, compared to the *hyper-exponential* distribution that has a coefficient of variation greater than one. Note that the exponential distribution has a coefficient of variation equal to one. The hypo-exponential has a minimum coefficient of variation of $1/k$, which corresponds to the Erlang distribution with shape parameter k and scale parameter μ (Wikipedia (2009)).

To compute its probability density function is not trivial. We give here a summary of the computations involved and refer the reader to Ross (1997). For the case of a series of two

exponentials, with rates $\mu_1 \neq \mu_2$, we have

$$\begin{aligned} f_{X_1+X_2}(t) &= \int_0^t f_{X_1}(s) f_{X_2}(t-s) ds \\ &= \int_0^t \mu_1 e^{-\mu_1 s} \mu_2 e^{-\mu_2 (t-s)} ds \\ &= \mu_1 \mu_2 e^{-\mu_2 t} \int_0^t e^{-(\mu_1-\mu_2)s} ds \\ &= \frac{\mu_1}{\mu_1 - \mu_2} \mu_2 e^{-\mu_2 t} + \frac{\mu_2}{\mu_2 - \mu_1} \mu_1 e^{-\mu_1 t}. \end{aligned}$$

Similar computations yield that, for a series of three exponentials,

$$f_{X_1+X_2+X_3}(t) = \sum_{i=1}^{3} \mu_i e^{-\mu_i t} \left(\prod_{j \neq i} \frac{\mu_j}{\mu_j - \mu_i} \right),$$

which suggests the general formula (for a series of k exponentials, with different μ_k's)

$$f_X(t) = \sum_{i=1}^{k} \prod_{j \neq i} \frac{\mu_j}{\mu_j - \mu_i} \mu_i e^{-\mu_i t}.$$

This can be proven by induction on k. Integrating on both sides from t to ∞ yields that the tail distribution function of X is given by

$$\mathbb{P}(X > t) = \sum_{i=1}^{k} \prod_{j \neq i} \frac{\mu_j}{\mu_j - \mu_i} e^{-\mu_i t}, \tag{2.1}$$

further denoted as $\bar{F}_k(t)$.

An intuitive derivation of the tail distribution of the hypo-exponential distribution is offered in Koole (2009).

Remark 1. If we have a system where abandonments are allowed and we assume exponential times (with rate γ) until abandonments occur, then abandonments can be incorporated in the Erlang delay model. The arrival rate is $\lambda(x, x+1) = \lambda$ and the departure rate for state $0 < x \leq s$ is equal to $\lambda(x, x-1) = x\mu$, where λ is the Poisson arrival rate to the system and μ is the service rate. However, the departure rate for higher states is different: $\lambda(s+x, s+x-1) = s\mu + x\gamma$, for all $x > 0$. Note that the system is always stable, independent of the values of λ, μ and s, as long as $\gamma > 0$. In this system the waiting time distribution, conditioned on the state, is not a gamma distribution anymore, as in the case of an $M/M/s$ queue, but a hypo-exponential distribution. For example, if a customer arrives in state $s+k$ (there are k waiting customers in front of him/her), then this customer has to wait the sum of exponentially distributed random variables with rates $s\mu + k\gamma$, $s\mu + (k-1)\gamma, \ldots,$ $s\mu$ before being served, which implies a hypo-exponentially distributed waiting time.

The proof of this result follows from the properties of the exponential distribution; for $h > 0$ small and $k > 1$, we have that

$$F_k(t+h) = \mu_k h F_{k-1}(t) + (1 - \mu_k h) F_k(t) + o(h).$$

Rewriting and taking the limit as $h \to 0$, the above becomes

$$\bar{F}_k'(t) = \mu_k(\bar{F}_{k-1}(t) - \bar{F}_k(t)), \text{ for } k > 1.$$

By differentiating Equation (2.1) and plugging it into the above formula, we conclude that, after some rewriting, Equation (2.1) gives the solution to $\bar{F}_k(t)$.

To derive the moments of the hypo-exponential distributions, a Laplace transform provides relatively easy computations. As a reminder, the Laplace transform of a non-negative random variable $X \geq 0$ with probability density function $\tilde{f}(x)$ is defined as

$$\tilde{f}(s) = \int_0^\infty e^{-st} f(t)dt = E[e^{-sX}] = \int_0^\infty e^{-st} dF(t).$$

In our case,

$$\tilde{f}(s) = \prod_{i=1}^{k} \frac{\mu_i}{\mu_i + s},$$

since we have a series of k exponentials with individual rates μ_i. One can compute the first two moments as follows:

$$\begin{aligned} E[X] &= -\tilde{f}'(0), \\ E[X^2] &= \tilde{f}''(0). \end{aligned} \tag{2.2}$$

It can be proven by induction that

$$\tilde{f}'(s) = -\tilde{f}(s) \sum_{i=1}^{k} \frac{1}{\mu_i + s}, \tag{2.3}$$

and

$$\tilde{f}''(s) = \tilde{f}(s) \left[\left(\sum_{i=1}^{k} \frac{1}{\mu_i + s} \right)^2 + \sum_{i=1}^{k} \left(\frac{1}{\mu_i + s} \right)^2 \right]. \tag{2.4}$$

By plugging (2.3) and (2.4) into (2.2) and working out the computations (with $\tilde{f}(0) = 1$), we find that

$$E[X] = \sum_{i=1}^{k} \frac{1}{\mu_i}, \tag{2.5}$$

$$E[X^2] = \left(\sum_{i=1}^{k} \frac{1}{\mu_i} \right)^2 + \sum_{i=1}^{k} \frac{1}{\mu_i^2}.$$

Hence it follows immediately that

$$Var[X] = \sum_{i=1}^{k} \frac{1}{\mu_i^2}. \tag{2.6}$$

2.4 Time Series Analysis

The analysis of experimental data, observed at different points in time, leads to new problems in statistical modeling due to correlations introduced by sampling adjacent points in time. Obviously, traditional statistical methods cannot be applied anymore, since they are based on the main assumption of independence between these observations. The newer approach that tries to solve this problem is commonly referred to as *time series analysis*.

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for the sampled data. A time series is defined as a collection of random variables indexed according to the order they are obtained in time. $\{x_t\}$ will further denote the value taken by the series at time t, with $t = 0, \pm 1, \pm 2, \ldots$. A simple example of time series is a collection of independent random variables, w_t, with mean 0 and finite variance σ_w^2. This time series is called *white noise*, due to its applications in engineering. A particular case of the white noise is the *Gaussian white noise*, for which w_t are independent normally distributed variables. If the stochastic behavior of time series could be explained in terms of the white noise, then classical statistical methods would suffice.

Auto-correlation and Cross-correlation

Two ways of introducing smoothness into time series models are moving averages and autoregressions. For example, if we have a white noise series and we replace the value w_t by the average of its current value and its immediate neighbors in the past and future, we obtain a smoothed series $\{v_t\}$:

$$v_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1}),$$

Now consider a time series denoted by $\{x_t\}$ and the white noise series. The output of equation

$$x_t = x_{t-1} - 0.9x_{t-2} + w_t,$$

computed successively for $t = 2, 3, \ldots, 500$, represents a smoothed series on x_t, built as a regression of the current value on the past two values of the series. The starting values x_{t_0}, x_{t_1} are known a priori. Since the regression is on values of the same series, the model bears the name of *autoregression*.

The dependence between two adjacent values x_s and x_t can be assessed through *covariance* and *correlation*. The *autocovariance* function measures the linear dependence between two points in the same series observed at different times and is defined as follows:

$$\gamma_x(s,t) = E[(x_s - \mu_{xs})(x_t - \mu_{xt})], \text{ for all } s \text{ and } t,$$

where μ_{xs} and μ_{xt} denote the mean over all possible events that could have produced x_s and x_t, respectively. Obviously, if $s = t$, the autocovariance reduces to the variance:

$$\gamma_x(t,t) = E[(x_t - \mu_{xt})^2].$$

The *autocorrelation* function (ACF) measures the linear predictability of x_t using the value x_s and is defined as

$$\rho_x(s,t) = \frac{\gamma_x(s,t)}{\sqrt{\gamma_x(s,s)\gamma_x(t,t)}}.$$

It can be easily seen that $-1 \leq \rho_x(s,t) \leq 1$, with ± 1 meaning a perfect linear prediction.

The autocorrelation not accounted for by lags $s+1$ to $t-1$ is called *partial autocorrelation* (PACF). In other words, the partial autocorrelation function measures the autocorrelation between x_s and x_t with the linear dependence of x_{s+1} through x_{t-1} removed.

We would also like to measure the predictability of another series y_t from the series x_t. We have the *cross-covariance* function

$$\gamma_{xy}(s,t) = E[(x_s - \mu_{xs})(y_t - \mu_{yt})]$$

and the *cross-correlation* function

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}}.$$

It is easily possible to extend the above to multivariate time series.

Although we have not made any special assumptions in the above definitions, a sort of regularity in the data behavior exists. This regularity is called *stationarity*. The *strict stationarity* identifies the equality

$$\mathbb{P}(x_{t_1} \leq c_1, \ldots, x_{t_k} \leq c_k) = \mathbb{P}(x_{t_1+h} \leq c_1, \ldots, x_{t_k+h} \leq c_k),$$

for all $k = 1, 2, \ldots$, time points $t_1, t_2, \ldots, t_k$, numbers $c_1, c_2, \ldots, c_k$ and all time shifts $h = 0, \pm 1, \pm 2, \ldots$. In other words, a series is strictly stationary if the probabilistic behavior of every collection of values $\{x_{t_1}, \ldots, x_{t_k}\}$ is identical to that of the shifted set $\{x_{t_1+h}, \ldots, x_{t_k+h}\}$.

This version of stationarity is too strong for most applications. Therefore, a milder version, called *weak stationarity* is commonly used. The conditions imposed on a time series x_t to be weakly stationary are:

- The mean value function μ_{xt} is constant and does not depend on t ($\mu_{xt} = \mu$ for all $t = 1, 2, \ldots$);
- The covariance function $\gamma_x(s,t)$ depends on s and t only through their difference $h = |s - t|$, called the lag.

Further on, the term *stationarity* will identify weak stationarity.

In this setting, for $s = t + h$, we have $\gamma_x(t+h, t) = \gamma_x(h, 0)$, which will be denoted by $\gamma_x(h)$. The autocorrelation function is then given by $\dfrac{\gamma_x(h)}{\gamma_x(0)}$ and the cross-correlation function by $\frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}$.

Usually, the analysis is performed on sampled data. For a sample of n data points, the above theoretical functions will be estimated by the following:

- The sample autocovariance function, defined as $\hat{\gamma}(h) = \dfrac{1}{n}\sum\limits_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})$, where $\bar{x}$ is the sample mean;
- The sample autocorrelation function, defined as $\hat{\rho}(h) = \dfrac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$;
- The sample cross-covariance function, defined as $\hat{\gamma}_{xy}(h) = \dfrac{1}{n}\sum\limits_{t=1}^{n-h}(x_{t+h} - \bar{x})(y_t - \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the sample means;

- The sample cross-correlation function, defined as $\hat{\rho}_{xy}(h) = \dfrac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}$.

The sample autocorrelation function has a sampling distribution that allows us to assess whether the data comes from a completely random series or whether correlations are statistically significant at some lags. The sample cross-correlation function can be examined graphically as a function of lag h to search for leading or lagging relations in the data.

Time Series Models

In the time series context, expressing x_t as a linear combination of previous values (e.g., $x_{t-1}, x_{t-2}, \ldots, x_p$) and lagged values of another series (e.g., $y_{t-1}, y_{t-2}, \ldots, y_{t-q}$) has a large range of applications. The simplest example of a regression model as such is the estimation of the trend within a data sample. Let x_t, for $t = 1, \ldots, n$, be a dependent time series and w_t the Gaussian white noise. Then, by fitting the model

$$x_t = \beta_1 + \beta_2 t + w_t,\ t = 1, \ldots, n,$$

we obtain the estimated coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ and we can estimate the trend. The model above assumes that the data has a stationary behavior around a trend. Therefore, estimating and eliminating the trend yields a stationary process.

A systematic class of models that handle time-correlated modeling and forecasting was developed by Box and Jenkins (Box and Jenkins (1970)). The autoregressive integrated moving average (ARIMA) models identify several components present in the data (e.g., trend, seasonal patterns), using sample ACF and sample PACF plots. The usual notation for this class of models is given below:

- The backshift operator B, defined as $B^k x_t = x_{t-k}$;
- Differences[1] of order d, ∇^d, defined as $\nabla^d = (1 - B)^d$;
- The autoregressive operator $\phi(B)$ of order p, defined as $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$;
- The moving average operator $\theta(B)$ of order q, defined as $\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$;
- The seasonal autoregressive operator $\Phi_P(B^s)$ of order P and seasonal lag s, defined as $\Phi_P(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{P_s}$, where $P_s = P * s$;
- The seasonal moving average operator $\Theta_Q(B^s)$ of order Q and seasonal lag s, defined as $\Theta_Q(B^s) = 1 + \Theta_1 B^s + \ldots + \Theta_Q B^{Q_s}$, where $Q_s = Q * s$;
- The seasonal difference ∇_s^D of order D, defined as $\nabla_s^D = (1 - B^s)^D$.

We can now introduce the ARMA, ARIMA, multiplicative seasonal ARMA and multiplicative seasonal ARIMA models.

A time series $\{x_t\}$ with $t = 0, \pm 1, \ldots$, is $\mathsf{ARMA}(p, q)$ if it is stationary and

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q},$$

[1] By differencing a time series, non-stationary components are eliminated and a stationary process is obtained.

with $\phi_p \neq 0, \theta_q \neq 0, \sigma_w^2 > 0$ and $\{x_t\}$ with a nonzero mean. In concise form, the model can be written as

$$\phi(B)x_t = \theta(B)w_t.$$

A time series $\{x_t\}$ with $t = 0, \pm 1, \ldots,$ is $\mathsf{ARIMA}(p, d, q)$ if $\nabla^d x_t = (1 - B)^d x_t$ is $\mathsf{ARMA}(p, q)$. In concise form, the model can be written as

$$\phi(B)\nabla^d x_t = \theta(B)w_t.$$

The multiplicative seasonal ARMA, denoted by $\mathsf{ARMA}(p, q) \times (P, Q)_s$, is given by

$$\Phi_P(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t.$$

The multiplicative seasonal ARIMA, denoted as $\mathsf{ARIMA}(p, d, q) \times (P, D, Q)_s$ is given by

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t.$$

To determine which type of model suits the data, a graphical analysis of the autocorrelation (ACF) and partial autocorrelation (PACF) functions is conducted. Table 2.1 and Table 2.2 show how to translate the conclusions from these graphs into a time series model. For graphical examples, see Chapter 3 of Shumway and Stoffer (1999).

	$AR(p)$	$MA(q)$	$\mathsf{ARMA}(p, q)$
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

Table 2.1: Behavior of ACF and PACF for ARMA models

	$AR(P)_s$	$MA(Q)_s$	$\mathsf{ARMA}(P, Q)_s$
ACF	Tails off at lags $ks, k = 1, 2, \ldots$	Cuts off after lag Qs	Tails off at lags ks
PACF	Cuts off after lag Ps	Tails off at lags $ks, k = 1, 2, \ldots$	Tails off at lags ks

Table 2.2: Behavior of ACF and PACF for ARIMA models

In practice it is usually the case that several models verify the conditions presented in the above tables. One of the methods to determine which model estimates the data behavior best is the Akaike's Information Criterion (AIC). It is the best method for regression models on small samples with a large number of parameters (see Chapter 2 from Shumway and Stoffer (1999)). The AIC is defined as

$$AIC = \ln \hat{\sigma}_k^2 + \frac{n + 2k}{n},$$

where n is the sample size, k the number of parameters and $\hat{\sigma}_k^2$ the maximum likelihood estimator for the variance, determined as

$$\hat{\sigma}_k^2 = \frac{RSS_k}{n},$$

with RSS_k the sum of squared residuals. The model that scores the highest AIC will explain the data behavior best.

For detailed information on time series analysis, we refer the reader to Shumway and Stoffer (1999).

2.5 Prediction Error Measurements

A crucial part of the modeling process is to evaluate of whether a given mathematical model describes the system accurately. This question can be difficult to answer as it involves several different types of evaluation. Usually, the easiest part of model evaluation is checking whether a model fits experimental measurements or other empirical data. A common approach to test a fit is to split the data into two disjoint subsets: training data and verification data. The training data are used to estimate the model parameters. An accurate model will closely match the verification data even though this data was not used to set the model's parameters. This practice is referred to as cross-validation in statistics.

How closely the model matches the verification data is measured by defining a metric to measure distances between observed and predicted data. The most often used metrics are the mean absolute percentage error (also known as MAPE), root mean squared deviation (RMSD) and variations of these, such as a *weighted* MAPE, *normalized* RMSD or coefficient of variation computed in terms of RMSD. We will further give an overview of the formulae involved and we refer the reader to Wikipedia (2009) for further details.

MAPE measures the accuracy of a fitted model. For n data points, we have

$$\mathsf{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|,$$

where A_t represents the observed value and F_t the forecasted value. The expression of the error relative to A_t error allows one to compare the error of fitted models that differ in level.

Although MAPE is very simple and convincing, the concept has a major drawback: if there are zero values among the observed data, a division by zero will occur. In order to avoid this problem other measures have been defined, for example the *weighted* MAPE (wMAPE), which is a ratio between the mean absolute error and the mean of the observed values:

$$\mathsf{wMAPE} = \frac{\frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|}{\frac{1}{n}\sum_{t=1}^{n}A_t} = \frac{\sum_{t=1}^{n}|A_t - F_t|}{\sum_{t=1}^{n}A_t}.$$

RMSD (also known as root mean squared error RMSE) is a frequently used measure of the amplitude of the residuals (the differences between values predicted by a model and the observed values):

$$\mathsf{RMSD} = \frac{1}{n}\sqrt{\sum_{t=1}^{n}(F_t - A_t)^2}.$$

The normalized root mean squared deviation (NRMSD or NRMSE) is the RMSD divided by the range of observed values:

$$\mathsf{NRMSD} = \frac{\mathsf{RMSD}}{x_{\max} - x_{\min}}.$$

The value is often expressed as a percentage, where lower values indicate less residual variance.

CVRMSD, or more commonly CVRMSE, is defined as the RMSD normalized to the mean of the observed values:

$$\mathsf{CVRMSD} = \frac{\mathsf{RMSD}}{\bar{x}}.$$

It represents the same concept as the coefficient of variation except that RMSD replaces the standard deviation.

Bibliography

[1] Box, G.E.P. and Jenkins, G.M. (1970). *Time Series Analysis, Forecasting and Control*. Holden-Day, 1st edition.

[2] Budge, S., Ingolfsson, A., and Zerom, D. (2008). Empirical analysis of ambulance travel times: the case of Calgary Emergency Medical Services. *Submitted to Management Science*. Available electronically at *http://www.business.ualberta.ca/aingolfsson/publications.htm*.

[3] Channouf, N., LEcuyer, P., Ingolfsson, A., and Avramidis, A.N. (2001). The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science*, **10**(1), 25–45.

[4] Erdoğan, G., Erkut, E., Ingolfsson, A., and Laporte, G. (2009). Scheduling ambulance crews for maximum coverage. *Journal of the Operational Research Society*. Advance online publication.

[5] Erkut, E., Ingolfsson, A., and Erdoğan, G. (2008). Ambulance location for maximum survival. *Naval Research Logistics*, **55**(1), 42–58.

[6] Erkut, E., Ingolfsson, A., Sim, T., and Erdoğan, G. (2009). Computational comparison of five maximal covering models for locating ambulances. *Geographical Analysis*, **41**(1), 43–65.

[7] Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, **5**, 79–14. Available at *www.math.vu.nl/ koole/articles/msom03/*.

[8] Henderson, S.G. and Mason, A.J. (2005). *Operations Research and Health Care - A Handbook of Methods and Applications*, chapter 4. Springer New York.

[9] Ingolfsson, A., Budge, S., and Erkut, E. (2008). Optimal ambulance location with random delays and travel times. *Health Care Management Science*, **11**(3), 262–274.

[10] Koole, G. (2006). Stochastic optimization. *Lecture notes*. Available electronically at *obp.math.vu.nl/edu/so/notes.pdf*.

[11] Koole, G. (2009). Optimization of business processes: An introduction to applied stochastic modeling. *Lecture notes*. Available electronically at *www.math.vu.nl/ koole/obp*.

[12] Restrepo, M. (2008). *Computational Methods for Static Allocation and Real-Time Redeployment of Ambulances*. Ph.D. thesis, Graduate School of Cornell University. Available electronically at *http://legacy.orie.cornell.edu/ shane/theses/MateoRestrpo.pdf*.

[13] Ross, S.M. (1997). *Introduction to Probability Models*. Academic Press, 6th edition.

[14] Shumway, R.H. and Stoffer, D.S. (1999). *Time Series Analysis and Its Applications: With R Examples*. Springer.

[15] Wikipedia (2009). Hypo-exponential distribution. Version from 23.07.2009, 18:00.

[16] Wikipedia (2009). Mean absolute percentage error. Version from 24.07.2009, 18:00.