Bus Inf Syst Eng 58(2):119–133 (2016)
DOI 10.1007/s12599-015-0391-3

RESEARCH PAPER

Data Analysis of Delays in Airline Networks

Lucian Ionescu · Claus Gwiggner · Natalia Kliewer

Received: 11 July 2013 / Accepted: 15 April 2015 / Published online: 25 June 2015

Abstract Cost-optimized airline resource schedules often imply a lack of delay tolerance in case of unforeseen disruptions, e.g. late check-ins, technical defects or airport and airspace congestion. Therefore, the consideration of timeliness and robustness has become an important topic in robust resource scheduling and a wide range of sophisticated scheduling approaches has been developed in recent years. However, these approaches depend on assumptions made concerning delay occurrences. A better understanding of delay mechanisms may lead to a better trade-off between cost-efficiency and robustness and is therefore the purpose of this paper. We provide a data-driven detection of decision rules for daytime delay trends, depending on spatio-temporal attributes. The focus is on interpretable rules whose prediction accuracy is compared to random forests as a non-parametric, automated modeling approach. The obtained results give an insight into both the nature of primary delay occurrence and the methodical potential of delay prediction in the context of robust resource scheduling.

Keywords Data mining · Data-driven delay analysis · Regression models · Robust airline resource scheduling

Accepted after five revisions by Prof. Dr. Suhl.

Electronic supplementary material The online version of this article (doi:10.1007/s12599-015-0391-3) contains supplementary material, which is available to authorized users.

Dipl.-Wirt.-Inf. L. Ionescu (✉) · Prof. Dr. N. Kliewer
Department of Information Systems, Freie Universität Berlin,
Garystr. 21, 14195 Berlin, Germany
e-mail: lucian.ionescu@fu-berlin.de

Prof. Dr. N. Kliewer
e-mail: natalia.kliewer@fu-berlin.de
URL: http://www.wiwiss.fu-berlin.de/kliewer/

Dr. C. Gwiggner
Institute for Operations Research, Universität Hamburg,
Von-Melle-Park 5, 20146 Hamburg, Germany
e-mail: claus.gwiggner@uni-hamburg.de
URL: http://www.uni-hamburg.de/OR

1 Introduction

At the day of operations airline transportation frequently has to deal with disruptions like technical breakdowns, late passengers, or bad weather conditions. These – mostly unforeseeable – events may cause resource allocation conflicts and thus schedule infeasibility for, e.g., crews and aircraft.

Resulting delay propagation necessitates recovery of schedules that imply high additional costs. In 2010, the recovery costs in Europe were estimated to exceed 1.25 billion Euros which are around 81 Euros per minute of delay – see Cook and Tanner (2011, p. 8) for further details. Although the number of flights in Europe decreased from 10 million in 2008 to 9.5 million in 2012, Eurocontrol expects an increase to 11.2 million flights in 2019 (Eurocontrol 2013, p. i). Thus, for airlines the consideration of schedule robustness has become an important topic in resource scheduling. The term of robustness involves the components stability and flexibility. Stability describes the degree of the ability of a schedule to remain feasible under changing operational environments. The main instrument to increase the degree of stability is the incorporation of buffer times between tasks. In contrast, flexibility means the degree in which a schedule can be adapted to changing environments, e.g., by simple and mostly cost-neutral opportunities to swap resources.

Unfortunately, an increasing degree of robustness comes along with an increase of the planned costs. Robust resource

scheduling approaches are efficient if a high increase of the robustness is gained with the planned costs only increasing slightly. However, the benefit of even highly efficient approaches depends on estimation of delays which are assumed to result from natural seasonal cycles and geographical patterns. Examples are holidays, differences between working days and weekend days, or varying weather conditions due to seasonal and geographical influences. Additional impacts result from varying demands which influences the flight schedule and network structure. In order to efficiently incorporate buffer times and resource swapping opportunities it is necessary to consider these cycles in long- and medium-term delay forecasts.

In the context of regular operations, we distinguish between primary and secondary delay. Delay that occurs due to exogenous disruptions is called primary delay. By contrast, secondary delay emerges from propagation effects in resource networks. It depends on scheduling decisions and can be avoided by robust scheduling. According to CODA (2011, p. 6), the ratio of secondary to primary delay has increased significantly from 0.54 in 2003 to 0.83 in 2008, meaning there were 0.83 min of secondary delay for 1 min of primary delay on average. As the latter depend on the network structure they can be influenced by scheduling decisions. For example, delay spreads through the flight network as a result of insufficient buffer times or missing cost-efficient recovery procedures. In particular, dependencies between different resource network layers for crews, aircraft and airport infrastructure, may lead to cascading propagation effects. For a survey on the impact of non-robust schedules see Atkinson et al. (2013).

There are two general approaches to deal with delay. The first one aims at increasing the robustness in regular daily operations, compensating delay resulting from ordinary disruptions like congestion effects, late check-ins or technical failure, to name but a few. The adaption to a changing environment should happen implicitly by delay absorption or by small manageable interventions. By contrast, there are highly competitive rescheduling approaches for catastrophic scenarios such as severe weather conditions, temporary airport closures or serious technical defects; see Clausen et al. (2010) for a recent survey. In these scenarios the main goal is to return to regular operations as quickly as possible. However, delay resulting from irregular massive disruptions cannot be anticipated in robust scheduling.

Referring to the first approach, the robustness of a schedule can be measured by the on-time performance, i.e., the sum of all delays. However, exogenous primary delay cannot be influenced by scheduling. Therefore, the relevant figure to consider is the secondary delay propagated due to insufficient buffer times between flights connected by the same resource. In consequence, a schedule A is more robust than a schedule B if the amount of propagated secondary delay is less in schedule A than in B. While secondary delay can be determined for example by simulating delay propagation, primary delay has to be generated independently. In consequence, the quality of robustness measurement depends on how realistically primary delay is modeled.

In this regard, the main goal of this paper is to examine the potential of data-driven delay modeling for robust resource scheduling. Therefore, we derive patterns in daytime trends of primary delay from historical data and evaluate their prediction accuracy by statistical modeling. Since resource scheduling is a long- and medium-term process, the focus of interest is on spatio-temporal variables that are available for delay prediction during the time horizon of scheduling; operational short-term predictor variables like weather conditions and congestion effects do not seem to be suitable in this context.

The study is embedded in a research project for robust airline resource scheduling, focusing on regular daily operations. According to Fink et al. (2014), one of the main challenges for model-based decision support is the necessity to take dynamic and stochastic system behavior into account when decisions are made. This data-driven research addresses the dynamic and stochastic counterparts of airline resource scheduling. In order to take into account the complexity of the data, we model the daytime trends in an approach following the idea of Analysis of Covariance (ANCOVA). The evaluation of the prediction accuracy of the observed patterns is performed by a model assessment step. The derived stochastic models and decision rules can be used to refine generators for primary delay in robust resource scheduling and the simulation of delay propagation. Note that resource scheduling is an airline-specific task and therefore all delay models are related to one airline only.

The remainder of this paper is organized as follows. In chapter 2 we discuss recent approaches to the usage of historical data for determining delay risks in robust airline resource scheduling. An overview on a generic resource scheduling framework is provided in order to clarify the contextual integration of our approach into the scheduling process. Finally, a discussion on statistical approaches for large data sets are presented. Chapter 3 gives a description of the available data and discusses problems in delay recording. In chapter 4, an exploratory data analysis is performed and decision rules concerning daytime trends in primary delay are derived. The prediction accuracy of the rules is evaluated by statistical model selection in chapter 5. Note that numerical results and interpretations in this study depend on the underlying data set. Nevertheless, the model assessment step is adaptable for the evaluation of varying delay trends. Furthermore, details on the model application in the related scheduling framework are

provided. Conclusions and directions for future research are addressed in chapter 6.

2 Delay Modeling in the Context of Robust Airline Resource Scheduling

Traditional airline resource scheduling deals with the minimization of planned costs. The usage of empirical delay information has become important for the field of robust resource scheduling. In this section we present recent advances with special regard to the usage of historical delay information. Furthermore, recent data mining approaches for large data sets are discussed.

2.1 Relevant Approaches for Delay Estimation in Robust Scheduling

Ageeva (2000) presents an approach to increase flexibility of schedules by incorporating swapping opportunities for aircraft. However, they do not consider the delay risk for incorporating swaps for flights that are likely to be disrupted. The evaluation of the approach is based on an increased number of swap opportunities which is considered as an indicator for increased flexibility.

The scheduled crew ground time is used as a deterministic indicator for stability in (Ehrgott and Ryan 2002, p. 141). Therefore, the difference between slack duration and expected duration of a departure delay, specified by flight routes, is used as a penalty factor for non-robustness. Weide et al. (2010) use a related measure for a heuristic iterative crew and aircraft scheduling approach. Schaefer et al. (2005) incorporate robustness by considering operational costs of crew pairing instead of planned costs. The operational costs are determined by separately simulated crew pairings in SimAir, a simulation framework that uses empirical delay distributions gained from historical data (Rosenberger et al. 2002, p. 373).

Yen and Birge (2006, p. 10) fit truncated gamma and log-normal distributions to real world data from Air New Zealand in order to generate disruption scenarios for a stochastic crew scheduling model. No information on the goodness-of-fit is given. Lan et al. (2006, p. 19) improve the stability of schedules by considering the delay propagation on aircraft routes. For the estimation they use historical data from the ASQP database. Gamma, log-normal and Weibull distributions are compared by means of classical goodness-of-fit tests. As a result, the log-normal distribution is found the best fit for 84 % of all flight arrival delays. The approach is also used by Dunbar et al. (2012). Note that both Yen and Birge (2006) and Lan et al. (2006) do not separately examine the possible impact of attributes such as time and location attributes of a flight in their delay models.

Tam (2011, pp. 89–121) also uses historical data for delay estimation. The flight delay is modeled by multiple-regression for every weekday. The regression terms consider the departure and arrival airport and the departure and arrival time. Note that no interactions between the variables are taken into account. The quality of the models is measured by R^2 only and the prediction error over an unknown data set has not been assessed. Dück et al. (2012, pp. 54–55) present a stochastic model for increasing the stability of crew and aircraft schedules. They use log-logistic and log-normal distributions per delay reason for the generation of primary delay scenarios. The expected delay of a flight is based on the convolution of different delay reasons. Delay due to weather, airspace and airport congestion are not considered in scheduling but in the subsequent simulation of generated schedules. Ionescu and Kliewer (2011) use the approach in a stochastic model for increasing the flexibility of crew schedules.

Shifting the focus away from robust scheduling, there is a variety of recent studies on the comprehension of delay mechanisms. Recent results include a large set of operational decision rules. Ball et al. (2007) give a survey on delay effects. In a statistical modeling approach for arrival delay, Hsiao and Hansen (2006) consider queuing, weather and seasonal effects. They discover negative daytime trends for queuing effects, i.e., delay occurring in the morning has a greater impact on delay propagation than in the evening. Their study extracts a large number of variables influencing delays which leads to a high explanatory power.

Xu et al. (2008) use regression models for estimating airport-related delay for the usage by operations control authorities. Again, besides the scheduled departure time and the scheduled turnaround time, especially short-term variables such as weather, operation demand in relation to airport capacity, ground holding and in-bound delays are considered. The prediction error is estimated by applying the model to an unknown test set. Tu et al. (2008) decompose delays into seasonal, propagation and random patterns. They concentrate on a specific airport in order to predict congestion delay effects. Wesonga et al. (2012) present the delay analysis of a single airport by using a variety of influential parameters like flight type, number of passengers, and weather conditions. Deshpande and Arıkan (2012) analyze the impact of scheduled block-times on the on-time arrival probabilities. Arıkan et al. (2013) aim at examining the impact of airline network structures and schedules on the reliability of the air-travel infrastructure. Therefore, they discuss stochastic models for actual block times, which follow a log-Laplace distribution. Secondly,

they develop a model for measuring the delay propagation through the flight network based on aircraft rotations.

The usage of historical data has become a standard procedure for estimating flight delay. However, we have to distinguish between micro- and macro-level delay estimation. For robust scheduling there is a demand for macro-level parameterization as the network characteristics have to be taken into account. Specific operational rules, e.g., for a single airport (Wesonga et al. 2012; Tu et al. 2008), cannot be used adequately for entire resource networks. The generalization of models with a large number of explanatory variables and resulting high prediction accuracy (e.g., Hsiao and Hansen 2006) for complete networks cannot be performed easily as it leads to an unmanageable model complexity. In the end, many short-term prediction variables with reasonable impact are not available during the long- and medium-term resource scheduling process.

As a consequence, delay estimation on a macro-level in robust scheduling is still a black box with mostly non-transparent assumptions. Most approaches are pared down to the determination of best-fitting distributions. The distinction of delay patterns for different parameters of flights are not taken into consideration. Only Ehrgott and Ryan (2002) use a distinction between flight routes but only consider average delay and standard deviation; Tam (2011) differentiates between time and location attributes. The necessity of our approach arises from this gap between operational delay studies and the requirements of robust scheduling approaches, implying a demand for statistical models that capture interpretable delay mechanisms on a macroscopic level. We provide the identification of systematic daytime trends in delay occurrence that are categorized by spatio-temporal attributes on the basis of related literature and practitioner's expertise. Derived decision rules are then analyzed with regard to their prediction accuracy. The comparison to automated model selection by a random forests approach is presented in order to examine the area of tension between prediction accuracy and interpretability of decision rules.

The resulting models and findings can be used as groundwork for a delay generator, enabling both resource scheduling and delay propagation simulation closer to reality. A generic framework for robust resource scheduling is illustrated in Fig. 1. For a given flight schedule, resource schedules (e.g., for crew and aircraft) are generated following certain scheduling strategies. Besides planning cost efficiency, robustness can be taken into account by considering primary delay and resulting propagation effects, see, e.g., Yen and Birge (2006) and Dück et al. (2012) for implementation details. The evaluation of robust scheduling strategies can be performed by means of event-based simulation. Whenever a delay occurs, it is either absorbed by buffer times or propagated to subsequent flights, depending on the propagation model. This approach only considers the stability of a schedule. Additionally, consideration of flexibility requires asks for recovery strategies in order to adapt the schedule to the current situation, see Shebalov and Klabjan (2006) or Ionescu and Kliewer (2011) for specifications. Besides primary delay, there are additional parameters influencing the schedule robustness, e.g., the network structure that determines the degree of freedom for scheduling decisions. Both hub-and-spoke and point-to-point network structures may contain flights that have to be flown in succession. Since flight schedules are predetermined in the context of this study, we assume network structures to be fixed. A lesser degree of freedom may also reduce the impact of refined primary delay models. Therefore, measuring the impact of network structures on the benefit of improved delay prediction will be part of future research.

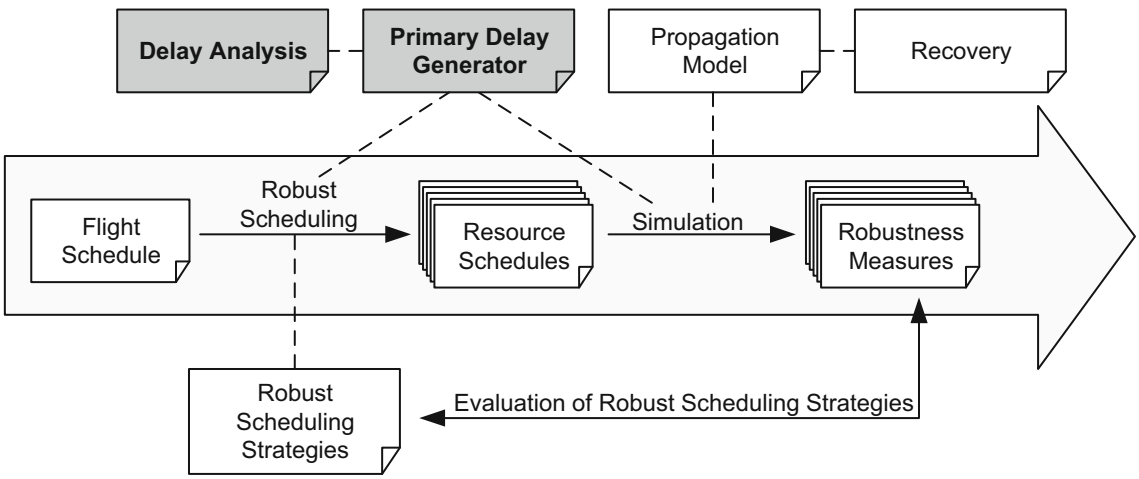

Fig. 1 Framework for evaluation of robust scheduling strategies

2.2 Data Mining Approaches for Large Data Sets

The development of prediction models and decision rules implies a field of tension between prediction accuracy and interpretability.

The *prediction accuracy* describes the relation between the model and the real data. High prediction accuracy means that there is a strong correlation between the predicted and the real value. In our context the usage of empirical distributions for delay predictions would have high prediction accuracy for short-term forecasts. However, this can lead to erroneous interpretations of the underlying mechanisms and result in wrong decision making.

An alternative approach is to focus on the delay generating mechanisms. This leads to the aspect of model *interpretability*. The benefit in this approach is the understanding of a substantial relationship between cause and effect. On the other hand, the prediction accuracy might be lower as only the most important patterns are captured.

Both targets of prediction accuracy interpretability are addressed in the analysis. An exploratory analysis is prepended to statistical modeling. The derivation of decision rules and the generation of predictive models are closely related to the field of data mining. Data Mining is often defined as the extraction of unexpected patterns in large data sets (Hand et al. 2001; Hastie et al. 2009). It uses statistical and algorithmic methods for descriptive and predictive problems.

Large data sets with thousands or millions of variables and observations pose challenges to formal statistical reasoning. For example, performing a large number of significance tests will reject by design a certain percentage of Null Hypotheses (e.g., Efron 2010). Moreover, with large sample sizes, standard errors of estimators tend to become so small, that even 'unimportant' differences between measured and *true* values are reported as significant. In predictive modeling, Big Data risks to favor complex models that 'mimic' the sample and its statistical fluctuation, but do not necessarily extract its underlying mechanisms (Hand et al. 2001, Chapt. 4.6.2; Hastie et al. 2009, Chapt. 7).

While for the purpose of short-term prediction some of these issues are resolved (for example by assessment of the bias-variance tradeoff), the data mining methodology does currently not provide a sound basis for the automated extraction of interpretable patterns (Breiman 2001a; Cox 2006; Cox and Wermuth 1996). As mentioned above, our strategy to avoid these pitfalls is to rely on descriptive methods, complemented by formal inferences, whenever possible.

3 Description of the Data

The following analysis is performed on a data set consisting of 2.5 million flight delay records provided by a major European airline for the time from March 2003 to February 2007. Only continental passage line flights are considered. Due to night flying restrictions there are only occasional flights between 10 p.m. and 6 a.m. which are excluded. Eventually, 2.2 million flight records are used for the analysis. Besides the number of passengers per flight, the available attributes can be differentiated into time, location and delay reason. The time aspect is characterized by scheduled and actual departure and arrival times in Central European Time (CET) stamps. For the determination of the local departure and arrival times we integrate information on time zones and daylight-saving time per airport, provided by openflight.org.[1]

The location is represented by the departure and arrival airport. The route attribute can be derived from origin-and-destination pairs. The network is based on a hub-and-spoke structure with two major hubs, where 38.7 % of all flights depart. 24.45 % of all flights are spoke-to-spoke connections. In addition to airports and routes, we take the network structure into consideration by distinguishing between the following directions: hub-to-spoke, spoke-to-hub and spoke-to-spoke.

A delay is defined as the nonnegative deviation between the scheduled and actual departure time. The departure time is defined as the time the aircraft leaves the gate. For every flight, up to four different departure delay reasons and their durations are recorded, based on standardized IATA Delay Codes. They define primary delays as exogenous effects with codes from 1 to 89, containing airline internal reasons, disruptions of the turnaround process, technical damages, or airport and airspace congestion, just to mention the main categories. The group of reactionary delays includes the codes from 90 to 96. These include waiting for passenger or load connections, for the late arrival of a resource such as aircraft or crew, and for decisions from operations control. Of course, the transition between endogenous and exogenous effects is fluent. The usage of the standardized IATA Delay Codes ensures the general adaptability of the approach.

Table 1 presents frequencies per number of departure delay records. Note that in this study we concentrate on positive delay values, early departures are declared to be on-time and thus set to a delay of 0 min. This is because negative delays do not propagate. In case of multiple records, secondary delay is mostly recorded first. Furthermore, only the delay is recorded that lead to late departures. Delay reasons that overlap in time are not entirely recorded, see Fig. 2 for illustration. Both of these effects lead to an underestimation of delay. In detail, 47.6 % of all

[1] OpenFlights Airports Database. http://sourceforge.net/p/openflights/code/757/tree/openflights/data/airports.dat. Accessed 27 October 2013.

Table 1 Occurrence frequencies of multiple delay records

# Delay records	0	1	2	3	4
Abs. frequency	911,568	1,043,622	236,148	17,184	1732
Rel. frequency	41.24 %	47.22 %	10.68 %	0.78 %	0.08 %

flights are primarily and 19.75 % secondarily delayed. 8.63 % of the secondarily delayed flights also contain primary delays.

Furthermore, primary delays as exogenous effects are inherently hard to predict. In most cases, primary delay can be recorded only if the departure (and arrival) times are effected. We also assume that existing patterns in primary delay occurrence are already taken into account by airlines in scheduling. Accordingly, as we will see in Sect. 5.3, the signal-to-noise ratio is rather low in the data.

Focusing on regular daily operations, we consider delays up to 180 min only, which cover 99.95 % of all flights, since larger delays imply serious disruptions that cannot be handled by regular robust scheduling. On the one hand the marginal costs for robustness become too large by taking into account such severe disruptions. On the other hand airlines have more effective capabilities to cope with these circumstances (Tam 2011, p. 91).

4 Exploratory Data Analysis

This section deals with an exploratory analysis of the data set. Beginning with a descriptive analysis, its aim is to provide a first overview of the data and to provide indications of cyclic patterns.

4.1 General Delay Trends

Figure 3 illustrates the average primary delay per day for the whole network. The delays seem to oscillate with a considerable amplitude during the entire span of time. The range of average delays is between 5 and 30 min. During winter months the average delay reaches its peak values more frequently. The black and dark grey lines are local estimations of a time trend using different smoothing parameter. This means that for a time point t, delays in its neighborhood are weighted in decreasing direction in order to determine its conditional mean. One can see that the line (1) is reasonably straight, indicating the absence of systematic changes of delay over time. Line (2) with a lower smoothing parameter indicates the peaks during winter months as mentioned above. Note that the derivation of time trends is a subjective decision, see (1990, p. 45) for further details.

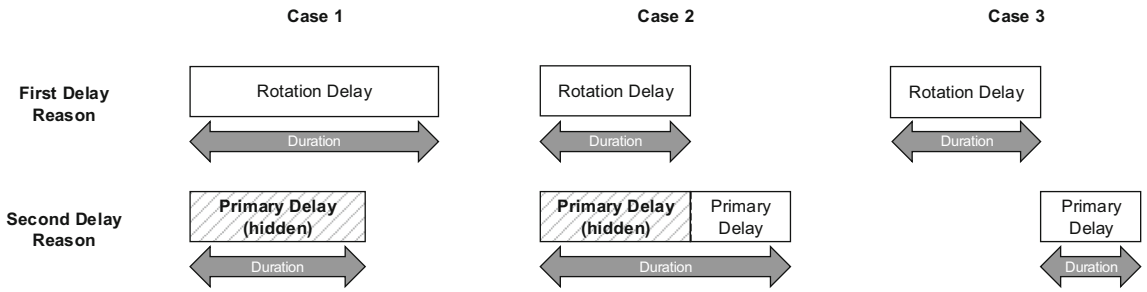

Fig. 2 Delay reasons overlapping in time

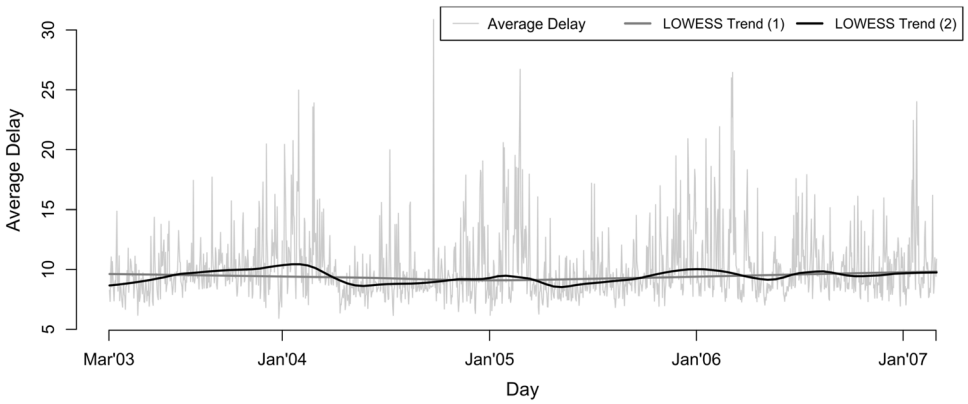

Fig. 3 Average primary delay per day over the time course

Table 2 Delay trends over the course of time

Period	1	2	3	4
Average	10.79	10.61	10.53	10.90
$\Delta(t_{i-1},t_i)$	–	−1.67 %	−0.75 %	3.51 %
Ratio	49.49 %	46.36 %	47.04 %	47.81 %
$d(t_{i-1},t_i)$	–	−6.32 %	1.47 %	1.64 %

Autocorrelation tests show that there are weak dependencies between Monday and Tuesday as well as between Thursday and Friday. We assume that this pattern can be explained by outward and return journeys at the beginning and the end of the working week.

Table 2 gives additional information on delay trends over time. In order to prevent zero-inflation we distinguish between the delay occurrence ratio and the length of delay. As the available data starts in March 2003, a period is defined as the interval from March to February of the following year each. Note that the relative differences are computed pairwise for succeeding periods. There are no obvious patterns. This is interesting, because one expects increasing delays due to a steady increasing flight demand. It seems that either slack capacity is available in the ground processes and the airspace system, or that the amount of resources grows with increasing demand. For further analysis, these patterns simplify the situation, since we can concentrate on seasonal effects in absence of complicating time trends and autocorrelations. But a straightforward derivation of delay occurrences from the flight schedule structure cannot be made.

4.2 Statistical Distributions for Description of Primary Delay Data

In the absence of strong autocorrelations and time trends, we empirically identify density functions that describe the delay during the different seasons. A first visual indication for well-fitting distributions is obtained by quantile-quantile-plots with a family of event-related distributions (Lindsey 2004, Chapt. 4). The log-normal, log-logistic and the Weibull turned out to be reasonable candidates. These distributions are fitted by Maximum Likelihood to the empirical data. The left panel of Fig. 4 illustrates an exemplary fit for summer months (May, Jun, Jul, and Aug). Log-normal and log-logistic seem to fit slightly better than Weibull. These results are consistent with (Lan et al. 2006, p. 19) who also consider these distributions as there are many small delays and only few very large delays. Taking the logarithm of the data leads to good fits with the normal and logistic distributions (right panel of Fig. 4). The logistic has a slightly better fit. However, there are crucial differences of both distributions for small values between 0 and 1. This can be explained by the data quality: only delay larger than one minute was considered, and the measurement unit is in minutes. Therefore, the logarithm for values smaller equal one will be distorted. The results can be reproduced for other data excerpts, i.e., not just for summer months.

Conventional χ^2-tests are not suitable for determining goodness-of-fit of theoretical distributions for large data as already small differences between observed and theoretical frequencies lead to a rejection of the null-hypothesis. For example, the null-hypothesis for the log-normal distribution is rejected on a 5 % level at values larger than 47.40 with 33° of freedom. Our sample statistics has a value of over 2000. This problem is already known from Berkson (1938).

4.3 Cyclic Patterns of Delay Occurrences

This section deals with determining patterns in primary delay. To this end, flights are categorized by temporal and network attributes. In detail, there are the attributes season, month and week, weekday and direction, for example from hub to spoke or the inverse. The local time, measured at each departure airport, is used to determine daytime trends. In the remainder of the analysis, hourly bins are used for the departure times, but smaller intervals showed similar results. The analysis intentionally focuses on the hub-and-spoke network structure and not on individual airports. In particular, the number of flights at individual spoke airports is so small that it is impossible to derive general daytime patterns. The same holds for an entirely route-based evaluation. Systematic dependencies between congestion indicators, such as the number of passengers and the primary delay length, cannot be observed. This is an indication that the airline has already eliminated predictable delay in its schedules.

Figure 5 exemplarily shows daytime trends in the different months (Jan to Dec from left to right) for spoke-to-hub flights at working days. Thinner lines in the background indicate the conditional average delay. The overlapping bold lines are the result of linear regressions. Note that the vertical axis depicts the logarithm of the actual delay.

In general, delay either grows or decreases during the day. Most months show a negative daytime trend as longer delays occur more often during morning hours. In contrast, in the summer months (May, Jun, Jul and Aug) a reverse daytime trend can be observed as evening hours display a higher average delay than morning hours. In conclusion, the daytime trend differs between months.

Systematical daytime trends can also be observed for different categories, i.e., for other flight directions and for the weekend. The daytime trends for these categories are

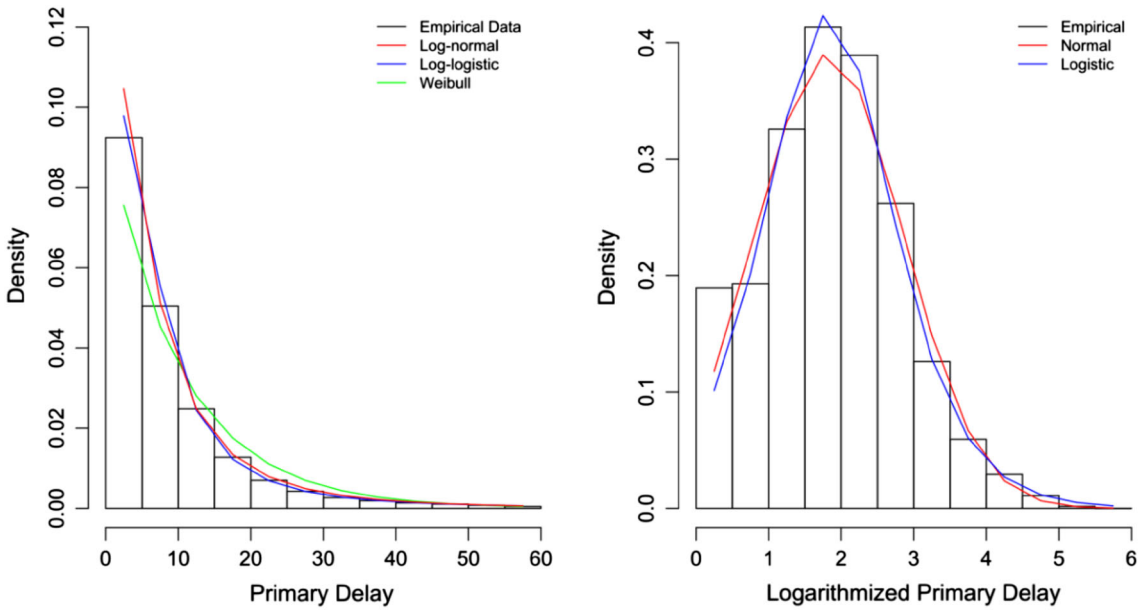

Fig. 4 Distribution for primary delays in summer months

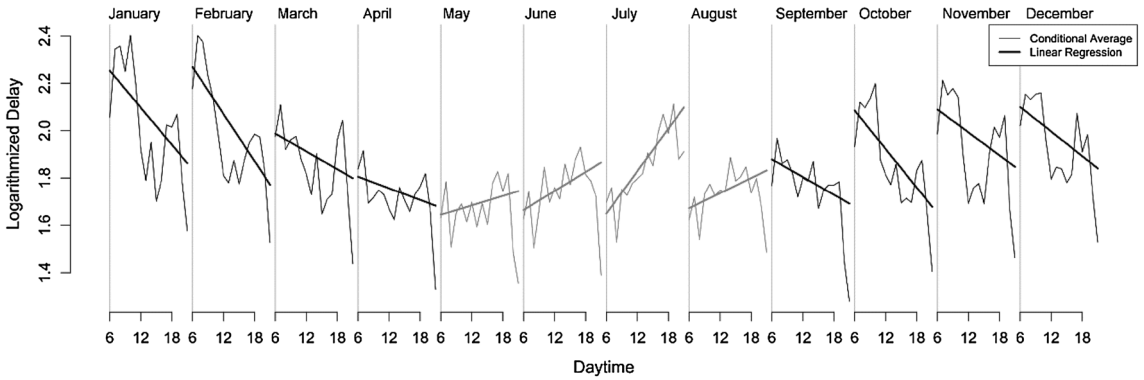

Fig. 5 Daytime trends per month for flights into hubs

similar with a slightly lower explanatory power. For the presentation of daytime patterns we use months as a seasonal attribute. The consideration of weeks instead of months is slightly more precise, especially for the location of transition points between winter and summer months. The summer cycle begins in week 17 (mostly end of April) and ends in week 36 (beginning of September). Without an exception, winter and summer weeks always follow their seasonal daytime trend.

Figure 6 illustrates spoke-to-spoke flights with the additional distinction between weekdays for the summer months and the rest of the year, respectively. The expected increasing daytime trend for summer is not valid for Monday and Saturday. All other weekdays, however, show the previously observed seasonal daytime trend. In the rest of the year, Friday and Sunday show a behavior that differs from the expected seasonal daytime trend.

Additionally it has to be said that for hub-to-spoke flights the dependency of weekdays can be observed, too – in the summer months there are negative daytime trends for Monday and Friday. However, the daytime trend for hub-to-spoke flights during the rest of the year is still slightly negative for Friday and Saturday, though it almost flattens out. Spoke-to-hub flights do not show a dependency on weekdays as their daytime trends follow the seasonal trend both in winter and summer.

We assume that there are peaks in the week structure that overlay the seasonal trends. On Monday morning and Friday evening there are peak values due to increased demands. By contrast, on Saturday evening very low

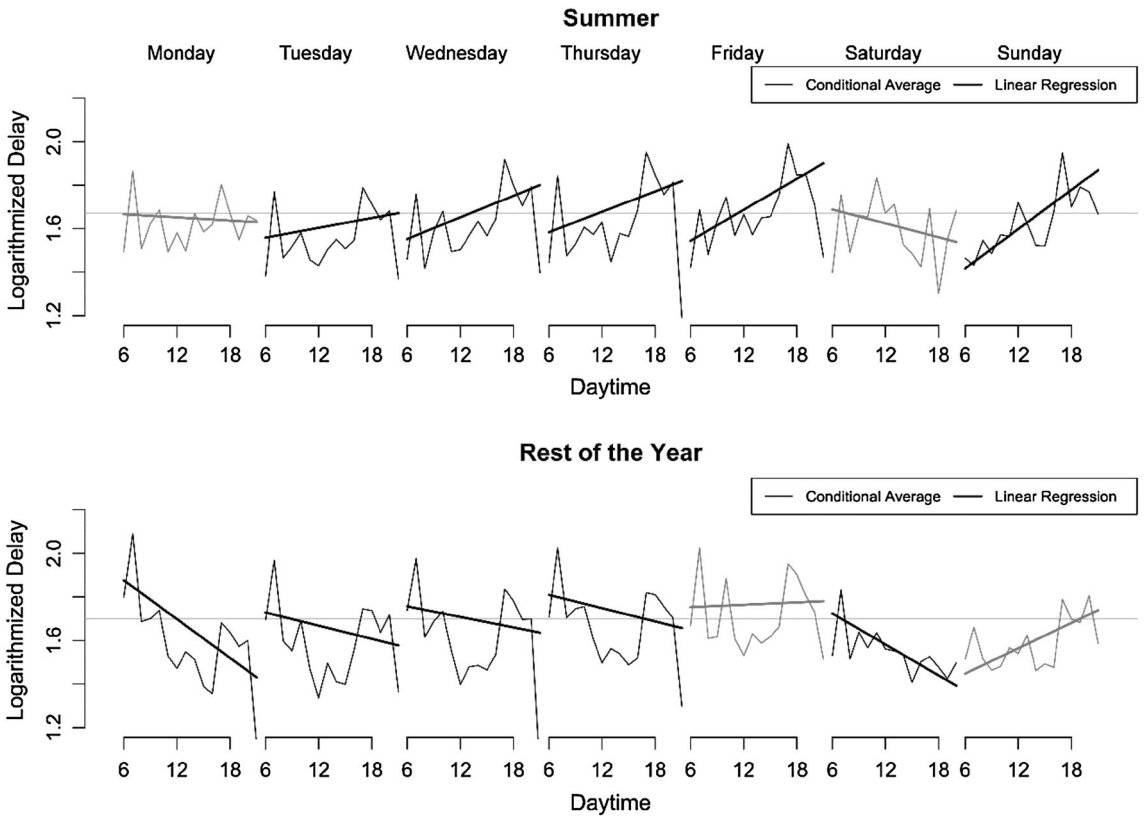

Fig. 6 Daytime trends per weekday for spoke-to-spoke connections

demand levels are expected. An important fact is that spoke-to-hub flights do not show these effects as they monotonously follow the daytime trend of the current season for every weekday. We suppose that this difference is an implication of the fact that the hub is not the final destination for most passengers.

Summarizing the above, the following decision rules can be derived from our exploratory analysis:

1. Regarding the average delay per hour, there is a positive daytime trend during the summer months (May, Jun, Jul, Aug), except on Mondays and Saturdays in case the arrival airport is a spoke.
2. By contrast, a negative daytime trend can be observed during the rest of the year, except on Fridays and Sundays for spoke-to-spoke connections.

We validate the rules for every single day. The first rule is valid for 57.86 % of all considered days, the second for 65.46 %, respectively. The weighted average for all days is 62.90 %. Interestingly, by taking into account seasons only, the error increases by just 2 %. The results align with results obtained by a CART analysis where the error remains almost constant when forcing additional splits beyond the seasonal one. This first evaluation resulting in poor validity of the rules strongly demands for a modeling approach that is capable of capturing the complexity of these mechanisms.

5 Model Selection and Assessment

In the previous sections we identified seasonal and monthly daytime trends that were positive in summer and negative during the rest of the year. We also discovered that on a daily level, these trends sometimes deviate from their seasonal component: Mondays and Saturdays during summer show a negative daytime trend, whereas Fridays and Sunday during winter show a positive daily trend. In this section we set up statistical models to quantify the predictive power of these findings. With statistical model we mean a model of the joint distribution of the observed data, along the common definitions such as (Cox 2006) or (Hastie et al. 2009).

More precisely, our problem is to model daytime trends in a number of spatio-temporal categories, such as flight directions, weekdays, and a seasonal component given by seasons, months or weeks. This is commonly referred to as ANCOVA (Analysis of Covariance). Two particularities of this approach are:

- *Daytime trends instead of average value*
  In an analysis of variance (ANOVA), the average value is estimated for each category. Categories with significantly different average values are identified by hypothesis tests. In our models, a daytime trend is fitted instead of the simple mean values. Such an analysis with a mix between categorical and continuous explanatory variables is called Analysis of Covariance (ANCOVA) (Lindsey 2004, p. 20).

- *Interactions*
  Classical ANCOVA models introduce different intercepts for each category only. This gives us for example one daytime trend for summer and a shifted one for winter. As identified in the previous chapter, these trends may differ in slope across the categories; summer trends are positive and for winter months they are negative. Such patterns can be modeled by *interactions* between the continuous and categorical covariates. Then, the prediction for time t in the k-th category is

  $$\mu(t) = \beta_{0k} + \beta_{1k} t$$

  corresponding to a linear model with intercept β_{0k} and slope β_{1k}. Instead of linear time-trends, we will later also fit cubic splines, corresponding to a basis expansion of the form $\mu(t) = \beta_{0k} + \sum_j \beta_{jk} f_j(t)$, where f_j is the j-th transformation of t. Note that the interaction between the k-th category and the time-variate is captured by the parameter β_{1k}.

The main assumptions in these models are their additive structure and their stochastic behavior. This means that for every category, the response is considered to be a random variable with mean being a function of time and constant variance. Instead of Gaussian variables, we assume log-normal variables. Other distributions, especially those belonging to the exponential family, are natural extensions to our approach. Note that we will not perform significance tests on estimated parameters, but only assess prediction accuracy of our models. Thus, distributional assumptions, including independence in the residuals, are not required at this stage of research (see, e.g., Weisberg (2005) for a discussion on stochastic assumptions in regression, and where they are needed). Auto-correlated data is traditionally modelled by time-series analysis. However, autocorrelation can sometimes already be explained by appropriate time-dependent covariates (Lindsey 2004, p. 10). Based on the descriptive analysis, on our purpose of the models and on the fact that we perform a macroscopic analysis, we believe that neglecting possible autocorrelation can be justified.

The remainder of this section is organized as follows: in a first step the predictive power of categorical variables of the identified daytime trends is determined. In order to maintain relative comparability, single models for the whole data set are considered. Subsequently, a residual analysis is performed and the prediction error on unknown data is estimated.

All model fitting is carried out in the R programming language (version 3.1.2) on an Intel i7 950 Quad-Core processor with 3.07 GHz and 24GB RAM. The most complex models can be estimated within several minutes; however, the bottleneck is the availability of memory.

5.1 Structural Model Selection

The first step deals with the selection of the model structure, and especially with the determination of the categorical variables. All models are fitted by maximum likelihood, more precisely by the iteratively reweighted least squares method. Table 3 illustrates the results. All models are based on delays on a log-scale. As a measure for the predictive quality of the models we penalize the likelihood by model complexity with the known information criteria AIC and BIC, see for example Hastie et al. (2009, pp. 230). The AIC is used to give the relative quality of different models on a given set of data. The differences of the AIC values are given as the differences to the previous model each, except for model S1, referring to L5.

We use the nominal parameters S (season), M (month), W (week), WE (weekday) and DI (direction) to determine the category. A linear regression is then fitted for the interaction between the categorical variables and the continuous variable D (daytime). The simplest models are a single daytime trend for all levels and categories (L1), and one daytime trend per season (L2). Already, AIC improves by 462 and 2038 units, respectively. Allowing for a

Table 3 Structural model selection

		Model	ΔAIC	ΔBIC	#Parameters
Linear models	L0	Mean value	–	–	1
	L1	D	−462	−450	2
	L2	D ° S	−2038	−2016	4
	L3	D ° M	−2355	−2126	24
	L4	D ° W	−3332	−2371	108
	L5	D ° S ° DI ° WE	−1031	−1305	84
	L6	D ° M ° DI ° WE	−5142	−338	504
	L7	D ° W ° DI ° WE	–	–	2268
Spline models	S1	D(3) ° S ° DI ° WE	−4614	−3927	168
	S2	D(5) ° S ° DI ° WE	−2779	−1819	252
	S3	D(15) ° S ° DI ° WE	–	–	672

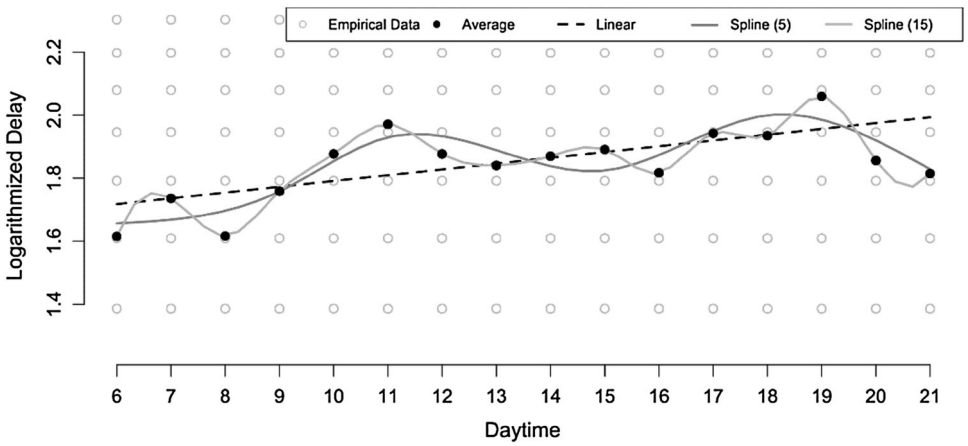

Fig. 7 Exemplary spline models for daytime trends of an exemplary category

daytime trend per month and week improves the AIC by 2355 (L3) and additional 3332 (L4) units. Our decision rule of the previous chapter, namely that daytime trends are permitted to differ among seasons, weekdays and directions, clearly improves the fit (L5, L6). Note that L6 already contains 504 regression parameters, and that model L7 cannot be computed anymore on the available system.

In order to further improve the prediction accuracy, we fit cubic splines (models S1 and S2) instead of linear trends (L5). Fitting regression splines is still linear in the parameters. What makes the difference, however, is a previous transformation of the continuous daytime variable according to a basis expansion (see Hastie et al. 2009, Chapt. 5.2). An example can be seen in Fig. 7 where we display the linear trend and two splines for a smaller data excerpt with the highest possible degree of freedom (15). The more degrees of freedom, the more splines are allowed. Numerical experiments showed us that the highest possible degree of freedom for model L5 is 5. The improvement in AIC and BIC is considerable, and higher values up to the maximum of 15 are in principle desirable. However, the resulting models can no longer be estimated or interpreted due to their complexity. Due to the same reason, the application of splines in models L6 and L7 is not possible.

In summary, the results confirm the observations of the exploratory analysis. All previously identified categorical variables lead to a considerable improvement of the prediction accuracy. Cubic splines improve the linear trends within the categories, although the most complex models can no longer be computed. However, this can be done in the following model assessment step. We concentrate on models represented by D(15) ° X ° DI ° WE, where X defines the seasonal component (season, month or week).

5.2 Residual Analysis

The main assumptions for our regression models were that for every category and every time point t, the data can be described by a log-normal distribution with the mean value as a cubic spline function of the time t and constant variance across the time.

If these assumptions are true, then the regression residuals, i.e., the differences between the predicted and observed values, have zero mean and follow a normal distribution with constant variance across time (normal because the logarithm of the delays was taken). A non-parametric bootstrap was performed to validate these assumptions. Figure 8 shows typical results for a category of the summer months, taking into account only flights into hubs. On the horizontal axis, the daytime is displayed in hourly slots. On the vertical axis, bootstrapped statistics of the residuals are displayed.

The black line depicts the residual averages. They follow a straight line on the zero value, thus the spline model does indeed capture the conditional mean of the data. The blue lines are pointwise estimates of the residual standard deviation at time t. If the homoscedasticity assumption is true, then they should be constant overtime. This is reasonably the case, although regarding a few time-points, e.g., at $t = 12$ or $t = 21$, care should be taken. Finally, the red lines depict 16 and 84 % quantiles of the residuals. They were selected according to those values which match the standard deviation of a normal distribution. Therefore, if the blue lines correspond to the red lines, the normality assumption is reasonable, at least to the second order. The 84 % quantiles (upper line) meet this condition very well. For the 16 % quantiles (lower line), differences to the standard deviation in the order of 10^{-1} are the rule, not the exception. This means that the model does not accurately predict delays that are smaller than the average delay at time t. This finding also corresponds to the poor fit of the normal distribution on the left tail in Fig. 4, although due to the regression function, no general relationship between marginal and residual distributions exist.

The quality of the bootstrap estimates was also assessed. It turned out that the standard errors of these estimates were

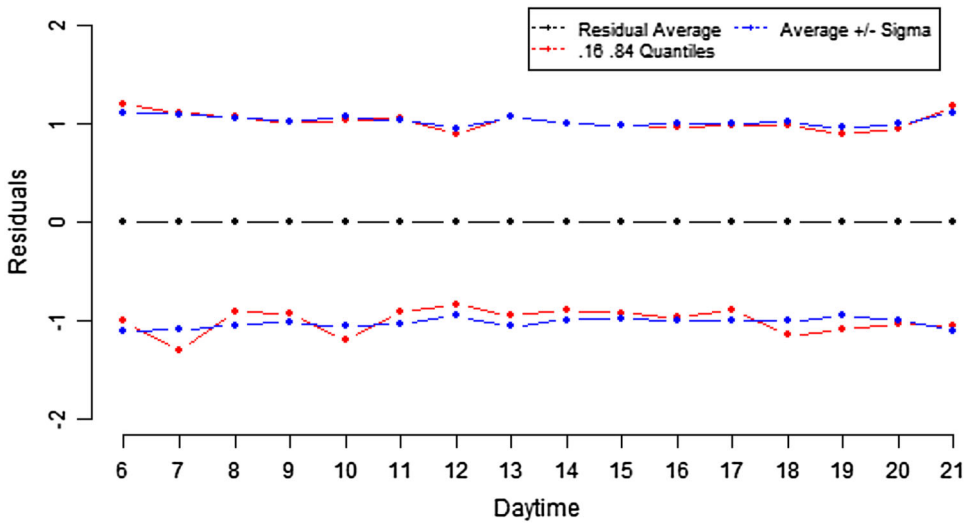

Fig. 8 Typical behavior of the residuals per daytime

in the order of 10^{-2} (see Figure A1 in the Appendix – available online via http://link.springer.com).

We conclude that the model assumptions are reasonably met for large delay predictions and require care for small delay predictions. Particularly for small categories, e.g., when splitting up the data by season, flight direction and weekdays, the effects of the left tail become apparent. Note that formal statistical inference about model parameters are not the target of this analysis, thus independence assumptions of the residuals are not verified.

5.3 Prediction Accuracy

While the structural model selection above is based on the analytical information criteria AIC and BIC, this section deals with a resampling technique to validate the prediction accuracy of the best structural models. As in this step the relative comparability is not of primary concern, the models can be split up into smaller and therefore less complex categorical models. The parameter estimations remain the same.

We follow the approach to model assessment, as described in Hastie et al. (2009, Chapt. 7). The target of this approach is the estimation of the expected extra-sample prediction error (EEPE), the prediction error that is independent of a given training data set. For comparison, the in-sample error (IE) for the training set is provided. In analogy with the the idea of ANOVA, the EEPE of our model m is computed as the residual sum of squares $RSS_{m,\tau}$ for each category concerning a validation set τ. It is compared to the $RSS_{a,\tau}$ of a model a that predicts the average value of each category, respectively. Then, the improvement factor

$$imp_{m,\tau}^{EEPE} = 1 - \left(\frac{RSS_{m,\tau}}{RSS_{a,\tau}}\right)$$

gives the amount of the variance in a validation set τ that can be explained by our model and thus the improvement of the EEPE obtained by model m. Concerning the IE, the computation of imp_m^{IE} follows analogously for the training set.

For the analysis we repeatedly split the data into training and validation data (70/30) for a large number of runs and estimate the corresponding test errors. In the end we average them for all categories, weighted by the respective number of flights. It turns out that with 100 runs, the averages converge towards a stable number. Table 4 shows the results for the models D(15) ° X ° DI ° WE, where X is one of the seasonal variables 'S', 'M' and 'W'. Other models from Sect. 5.1 are dominated by these models in both IE and EPEE.

On average, the EEPE can be improved by 1.95 % for model E1. In this model, 7.5 % of all categories show an EPEE improvement of more than 3 %. One can also observe the bias-variance-tradeoff by means of IE and EEPE. While the best in-sample error can be achieved by choosing more categories (E3), these models perform quite poorly regarding the EEPE. Furthermore, some categories in E2 and E3 have a negative EEPE value, especially in categories containing a small number of flights. In particular, these are the categories for spoke-to-spoke flights and during the weekend. It turns out that this effect is merely associated with the small category sizes.

For an illustration, see Fig. 9 which exemplarily shows the relative improvement of the EPEE under the squared error for model E2. The horizontal-axis shows the number of flights within the category concerned. As we consider flights that are delayed at least by 1 min, it is obvious that categories with hub-to-spoke flights are larger than those with spoke-to-hub flights, as the former are more likely to be delayed. Due to the hub-and-spoke structure there are less direct connections between spoke airports.

Regarding the IE, the results are comparable with the ones obtained by non-parametric random forests that

Table 4 Model evaluation

Model		Category			Improvement		
		#Categories	Ø #Flights	EEPE <0 (%)	IE (%)	EEPE (%)	Abs
E1	D(15) ° S ° DI ° WE	42	23,365	–	2.14	**1.95**	0.964
E2	D(15) ° M ° DI ° WE	252	3,894	1.50	3.02	1.93	0.971
E3	D(15) ° W° DI ° WE	1134	865	7.10	**5.73**	1.23	0.984

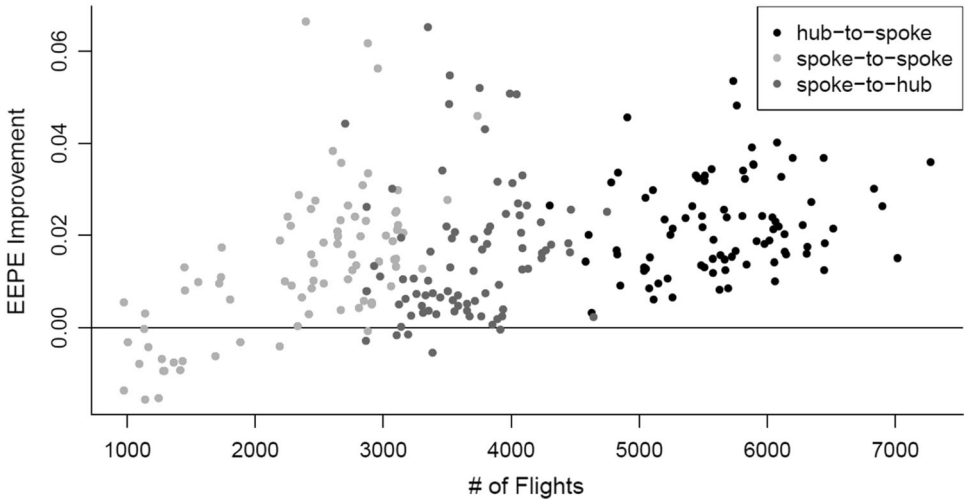

Fig. 9 Improvement per category for model D (15)° M° DI° WE

independently grow a large number of regression trees by repeated bootstrap sampling of the data (Breiman 2001b). In our case, random forests are applied independently from our previous decision rules. However, they cannot be applied to the whole data set due to memory restrictions, and repeated sub-sampling does not provide reasonable results. The same holds for the seasonal model (E1). An application based on monthly (E2) and weekly (E3) basis results in an IE of 3.71 and 6.39 %, respectively. It becomes apparent that these results are slightly better but still in the order of the ones obtained by our modeling approach. Prediction accuracy estimation for unknown data is performed internally in random forests by out-of-bag sampling (Hastie et al. 2009, pp. 592). For the prediction of x_i trees are used that do not contain the observation for x_i in their bootstrap sample used for growth. The prediction accuracy can be increased by 3.20 % (E2) and 4.13 % (E3), respectively. These values are expected since, in contrast to our decision rules, random forests implicitly provide an individual dynamic rule selection for all categories – leading to higher prediction accuracy in conjunction with a lack of interpretability.

Finally, the absolute deviation between observed and predicted delay is a metric that is easy to interprete, since its unit is in minutes. While the average absolute deviation of the mean model L0 is 8.29 min, it can be reduced by nearly 1 min to 7.32 by using the best ANCOVA-model E1. The best categories even show an improvement of the average absolute prediction error of 2.5 min (see Figure A2 in the appendix). These results are based on the presented daytime trends only. The absolute improvement can be used for estimating the benefit in real costs by linking them to specific airline's delay cost rates.

5.4 Model Application

Finally, we describe the incorporation of the generated models into the framework for robust resource scheduling and delay propagation simulation for resource schedules. For a given set of flights F in a schedule, it is necessary to determine the predicted primary departure delay d_f for each flight $f \in F$. All flights are part of a category that – in our resulting models – is currently determined by the direction, a seasonal component and the weekday.

For every category a regression model is fitted for the daytime variable.[2] For a given daytime t, the expected primary delay $\hat{\mu}(t)$ and its standard deviation $\hat{\sigma}(t)$, depending on the departure time t, is determined per category. Since we use log-scale delays, a random number $X \sim N\left(\hat{\mu}(t);\ \hat{\sigma}(t)^2\right)$ can then be picked from the normal

[2] The parameters of an exemplary cubic spline model are given in Table A1 in the appendix (available online via http://link.springer.com).

distribution. Finally, the resulting primary delay d is computed as $d = e^X$.

Concerning scenario-based robust resource scheduling as in (Yen and Birge 2006) or (Dück et al. 2012), the robustness evaluation of a resource schedule takes into account a set of primary delay scenarios Ω. With a given set of flights F, a delay scenario $\omega \in \Omega_F$ represents random variables for primary delay that result in deviations from departure times of the flights $f \in F$. Now, the primary departure delay of flight f in scenario ω is $d_{f\omega}$ which can be drawn analogously from our model.

With primary delays given, the robustness of resource schedules can be measured by determining the amount of propagated delay. An exemplary delay propagation model for aircraft and crew that is suitable for our approach is provided by Dück et al. (2012).

6 Summary and Outlook

During airline operations, exogenous disruptions often lead to delay that may result in infeasible resource schedules. The estimation of delay based on historical data is a recent topic in robust airline scheduling.

A better understanding of delay mechanisms may lead to a better trade-off between cost-efficiency and robustness and is therefore the purpose of this paper. We provide a regression modeling approach for daytime delay trends based on a data-driven detection of spatio-temporal patterns. The focus is on interpretable rules whose prediction accuracy is compared to random forests as a non-parametric, automated modeling approach.

First, decision rules were derived that describe daytime delay trends in spatio-temporal categories. For example, there is a positive daytime trend during summer, except on Mondays and Saturdays when the arrival airport is a spoke. Thus, we can state that the daytime trend depends on the interaction of the considered attributes. In order to validate these rules, we performed a quantitative evaluation by means of statistical modeling. From a technical point of view, the nature of our problem is related to the analysis of covariance (ANCOVA). The highest prediction accuracy so far can be achieved by spline models for daytime trends, taking into account interactions between the categorical variables season, direction and weekday. Although the derived decision rules, taken as a whole, are valid for only 62.90 % of all days, this leads to a reduction of the absolute prediction error by about 1 min on average. In particular categories, our approach leads to an even higher improvement of the prediction accuracy. The overall prediction accuracy is comparable to non-parametric random forests that imply an individual categorization and rule selection but lack interpretability.

However, we can assume that in general, primary delays are inherently hard to predict in the long-term on a macroscopic level. In this context, one always has to take into account that delay recording underlies constraints that lead to underestimation, e.g., predictable delay may already be prevented by scheduling decisions of an airline. In close connection to this, it is desirable to check to which extent the findings may be generalized regarding other airline delay data.

A lesson learned during this research was the discovery of the low signal-to-noise ratio of the time trends. They look promising on aggregated data, asking for further investigation and interpretation. During the statistical analysis, the variance of the delays around these time trends became apparent. Methodologically interesting was that, due to the large data sets, the standard errors of statistical estimators were so small that the resulting inferences were no longer conclusive. This is a general challenge of data-driven approaches that aim to argue by other means than predictive accuracy.

Future work shall therefore identify the conditions, under which accurate predictions of primary delay are feasible. The generated prediction models can then be implemented into a scheduling and simulation framework in order to obtain a more realistic evaluation of schedule robustness. The emerging question is to what extent an improved delay prediction affects the buffer management in hub-and-spoke networks and whether it actually leads to significant improvements of the robustness of schedules.

Acknowledgments This research was supported by a grant from the German Research Foundation (DFG, Grant No. KL2152/3-1). The authors are grateful to the anonymous referees for their helpful advice and feedback.

References

Ageeva Y (2000) Approaches to incorporating robustness into airline scheduling, Master's thesis, Massachusetts Institute of Technology

Arıkan M, Deshpande V, Sohoni M (2013) Building reliable air-travel infrastructure using stochastic models of airline networks. Oper Res 61(1):45–64

Atkinson S, Ramdas K, Williams JW (2013) The costs of inefficient robust scheduling practices in the U.S. airline industry. http://faculty.london.edu/kramdas/Cost%20of%20inefficient%20airline%20robust%20scheduling%20-%202013_04_16.pdf. Accessed 20 April 2015

Ball MO, Barnhart C, Nemhauser G, Odoni A (2007) Air transportation: irregular operations and control. Trans Handb Oper Res Manag Sci 14:1–73

Berkson J (1938) Some difficulties of interpretation encountered in the application of the Chi square test. J Am Stat Assoc 33(203):526–536
Breiman L (2001a) Statistical modeling: the two cultures. Stat Sci 16(3):199–231
Breiman L (2001b) Random forests. Mach Learn 45(1):5–32
Clausen J, Larsen A, Larsen J, Rezanova NJ (2010) Disruption management in the airline industry – concepts, models and methods. Comput Oper Res 37(5):809–821
CODA (2011) Planning for delay: influence of flight scheduling on airline punctuality. Eurocontrol Trends Air Traffic 7
Cook AJ, Tanner G (2011) European airline delay cost reference values. University of Westminster, London
Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge
Cox DR, Wermuth N (1996) Multivariate dependencies: models, analysis and interpretation. CRC Press, New York
Deshpande V, Arıkan M (2012) The impact of airline flight schedules on flight delays. Manuf Serv Oper Manag 14(3):423–440
Diggle PJ (1990) Time series: a biostatistical introduction. Oxford University Press, Oxford
Dück V, Ionescu L, Kliewer N, Suhl L (2012) Increasing stability of crew and aircraft schedules. Transp Res C 20(1):47–61
Dunbar M, Froyland G, Wu CL (2012) Robust airline schedule planning: minimizing propagated delay in an integrated routing and crewing framework. Transp Sci 46(2):204–216
Efron B (2010) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press, Cambridge
Ehrgott M, Ryan DM (2002) Constructing robust crew schedules with bicriteria optimization. J Multi Criteria Decis Anal 11:139–150
Eurocontrol (2013) Eurocontrol seven-year forecast February 2013 – Flight movements and service units 2013–2019. Brussels
Fink A, Kliewer N, Mattfeld D, Mönch L, Rothlauf F, Schryen G, Suhl L, Voss S (2014) Model-based decision support in manufacturing and service networks. Bus Inf Syst Eng 6(1):17–24
Hand D, Mannila H, Smyth P (2001) Principles of data mining. MIT Press, Cambridge
Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning. Springer, New York
Hsiao CY, Hansen M (2006) Econometric analysis of US airline flight delays with time-of-day effects. J Transp Res Board 1951(1):104–112
Ionescu L, Kliewer N (2011) Increasing flexibility of airline crew schedules. Procedia Soc Behav Sci 20:1019–1028
Lan S, Clarke JP, Barnhart C (2006) Planning for robust airline operations: optimizing aircraft routings and flight departure times to minimize passenger disruptions. Transp Sci 40(1):15–28
Lindsey JK (2004) Statistical analysis of stochastic processes in time (14). Cambridge University Press, Cambridge
Rosenberger JM, Schaefer AJ, Goldsman D, Johnson EL, Kleywegt AJ, Nemhauser GL (2002) A stochastic model of airline operations. Transp Sci 36(4):357–377
Schaefer AJ, Johnson EL, Kleywegt AJ, Nemhauser GL (2005) Airline crew scheduling under uncertainty. Transp Sci 39(3):340–348
Shebalov S, Klabjan D (2006) Robust airline crew pairing: move-up crews. Transp Sci 40(3):300–312
Tam B (2011) Optimisation approaches for robust airline crew scheduling. PhD Thesis, School of Engineering, University of Auckland
Tu Y, Ball MO, Jank WS (2008) Estimating flight departure delay distributions – a statistical approach with long-term trend and short-term pattern. J Am Stat Assoc 103(481):112–125
Weide O, Ryan D, Ehrgott M (2010) An iterative approach to robust and integrated aircraft routing and crew scheduling. Comput Oper Res 37(5):833–844
Weisberg S (2005) Applied linear regression. Wiley, New York
Wesonga R, Nabugoomu F, Jehopio P (2012) Parameterized framework for the analysis of probabilities of aircraft delay at an airport. J Air Transp Manag 23:1–4
Xu N, Sherry L, Laskey KB (2008) Multifactor model for predicting delays at us airports. Trans Res Rec J Transp Res Board 2052(1):62–71
Yen JW, Birge JR (2006) A stochastic programming approach to the airline crew scheduling problem. Transp Sci 40(1):3–147