Identifying and Responding to Outlier Demand in Revenue Management

Nicola RENNIE[a,*], Catherine CLEOPHAS[b], Adam M. SYKULSKI[c], and Florian DOST[d,e]

[a]STOR-i Centre for Doctoral Training, Lancaster University, LA1 4YW, UK. (n.rennie@lancaster.ac.uk)

[b]Institute for Business, Christian-Albrechts-University Kiel, Kiel, Germany. (cleophas@bwl.uni-kiel.de)

[c]Dept. of Mathematics and Statistics, Lancaster University, LA1 4YF, UK. (a.sykulski@lancaster.ac.uk)

[d]Management Science and Marketing Division, Alliance Manchester Business School, M15 6PB, UK.

[e]Institute of Business and Economics, Brandenburg University of Technology, 03046 Cottbus, Germany.
(florian.dost@b-tu.de)

[*]Corresponding Author: Nicola Rennie (n.rennie@lancaster.ac.uk)

Abstract

Revenue management strongly relies on accurate forecasts. Thus, when extraordinary events cause outlier demand, revenue management systems need to recognise this and adapt both forecast and controls. Many passenger transport service providers, such as railways and airlines, control the sale of tickets through revenue management. State-of-the-art systems in these industries rely on analyst expertise to identify outlier demand both online (within the booking horizon) and offline (in hindsight). So far, little research focuses on automating and evaluating the detection of outlier demand in this context. To remedy this, we propose a novel approach, which detects outliers using functional data analysis in combination with time series extrapolation. We evaluate the approach in a simulation framework, which generates outliers by varying the demand model. The results show that functional outlier detection yields better detection rates than alternative approaches for both online and offline analyses. Depending on the category of outliers, extrapolation further increases online detection performance. We also apply the procedure to a set of empirical data to demonstrate its practical implications. By evaluating the full feedback-driven system of forecast and optimisation, we generate insight on the asymmetric effects of positive and negative demand outliers. We show that identifying instances of outlier demand and adjusting the forecast in a timely fashion substantially increases revenue compared to what is earned when ignoring outliers.

Keywords: Revenue management; Simulation; Forecasting; Outlier detection; Functional data analysis.

1   Introduction

In the last 40 years, *revenue management (RM)* has become an indispensable business practice, particularly for transport service providers such as airlines and railways (Weatherford, 2016). RM solves an optimisation problem, where firms decide on offers for perishable products, usually with the objective of maximising revenue. This optimisation assumes a fixed capacity, low marginal cost, and a given *demand forecast*. In that regard, Weatherford and Belobaba (2002) highlight that inaccurate demand forecasts can significantly diminish the achieved revenue. Banerjee et al. (2019) point out that detailed demand forecasts also support in further planning steps, such as network resource and fuel planning.

Cleophas et al. (2017) list several causes for *forecast inaccuracies*: on the one hand, the unavoidable variance of day-to-day demand prohibits perfectly accurate forecasts. On the other hand, any flaw in the forecast model, including both the predictive time series component and the customer choice model naturally causes model-based forecast errors. Finally, sudden shifts in the market may cause short-term, temporal *outliers*. For example, when the system does not account for special events such as a sports championship or a trade fair, these will cause observed demand to systematically deviate from predictions.

We focus on such *demand outliers* in the domain of revenue management for passenger transport, specifically railways and airlines. In this domain, RM via capacity controls optimises *booking limits*, which specify the number of units that can be sold per fare class and time in a fixed *booking horizon*. Accordingly, sold units are also termed *bookings*. The distribution of bookings over intervals of the booking horizon constitutes a *booking pattern*. Booking patterns may be aggregated across fare classes and are reported either for single resources, such as flight legs, or for complementary combinations of resources, such as network itineraries. Here, we focus on aggregated booking patterns as reported for single resources, such as a single flight or a railway connection.

Common RM demand forecasting techniques estimate demand from historical booking patterns and booking limits (Weatherford, 2016). Accordingly, we let outlier detection rely on the same data. We follow the definition by Hawkins (1980) and define an outlier as 'an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.' Detection can either apply *online*, within the booking horizon and considering partial booking patterns, or *offline*, after a booking horizon, when the complete pattern can be analysed. Demand outliers affect revenue

management systems in two ways: (i) in *foresight*, the flawed forecast results in non-optimal capacity allocations; and (ii) in *hindsight*, the outlier can contaminate the data that underlies future forecasts. Accordingly, online detection can improve foresight, whereas offline detection can improve hindsight. To detect outliers, functional data analysis, where each booking pattern is treated as an observation of a function over time, is a natural place to turn to. Functional approaches can detect outliers in both magnitude and shape of an observed booking pattern. In other words, they can detect outliers that deviate across the entire booking horizon and those that deviate in only part of the booking horizon. Effective detection in online and offline settings has to be capable of identifying both types of outliers.

By investigating practical RM implementations in the airline and railway industry, we find that the current process relies on analysts, who manually examine booking patterns. When analysts perceive demand outliers, they attempt to compensate by adjusting the reported data, the forecast, or the booking limits. The decision of whether an adjustment is necessary and in what form depends on the analysts' intuition. As noted by Cleophas et al. (2017) and Banerjee et al. (2019), little existing work systematically measures the effect of such interventions. There is even less consideration of providing systematic analytics support for the related decisions. However, research on human decision making in general, and judgemental forecasting in particular, clearly demonstrates fallibility and bias (O'Connor et al., 1993; Lawrence et al., 2000, 2006). This motivates the need for automated alerts to highlight outliers and thereby support analysts.

To our knowledge, we are the first to propose an automated methodology for outlier detection in the RM domain. Specifically, this paper makes the following contributions: (i) proposing a novel outlier detection approach, combining functional data analysis and time series extrapolation, which improves overall detection performance; (ii) providing a simulation-based framework for generating regular and outlier booking patterns, and evaluating their effect throughout the RM process; (iii) demonstrating the asymmetric effects of outliers on RM performance; (iv) quantifying the benefits from successful online or offline outlier detection for RM; and (v) demonstrating the use of such outlier detection in an application to empirical railway booking data.

2 RM Forecasts and Forecast Evaluation

The importance of accurate forecasts as input to revenue optimisation is well-documented in the literature. Authors are largely concerned with forecasting customer demand (Pereira (2016), Weatherford and Pölt (2002), Talluri and Van Ryzin (2004)), although forecasting cancellations and no-shows has also been explored (Morales and Wang, 2010). Weatherford and Belobaba (2002) confirm previous findings that inaccurate demand estimates can significantly impact revenue. Under the use of optimisation heuristics such as Expected Marginal Seat Revenue (EMSRb) (Belobaba, 1989), under- or over-forecasting can even be beneficial. As described by Mukhopadhyay et al. (2007), most RM systems require forecasts of the *actual* demand, rather than the *observed* demand. The actual demand consists of both observed demand and customer requests that were denied due to restrictive booking limits. Actual demand is difficult to observe in practice, and so must be estimated. To this end, Weatherford and Pölt (2002) survey various techniques.

When allowing for inaccurate demand forecasts, much RM research focuses on rendering the optimisation component more robust or forecast-independent, as detailed in the contributions reviewed in Gönsch (2017). In another review, Cleophas et al. (2017) point out that there is little research into the effects of manually adjusted forecasts in RM. Mukhopadhyay et al. (2007) propose a method for measuring the performance of adjusted and unadjusted forecasts. They find that if analysts can reliably improve demand forecasts on critical flights, significantly more revenue can be generated. Zeni (2003) describe a study at US Airways, which aimed to isolate and estimate the value of analyst interactions. According to that study, around 3% of the additional revenue generated within the duration of the study could be attributed to analyst input.

Given that experiments in a live RM system carry significant risks, the use of simulation for evaluation is common. Additionally, simulation studies enable *a priori* knowledge about the true demand generation process, which can never be known in a real-world setting. Frank et al. (2008) discuss the use of simulation for RM and provide guidelines; in a related effort, Kimms and Müller-Bungart (2007) consider demand modelling for RM simulations. The paper at hand follows these contributions in establishing a simulation-based framework to generate outlier observations. Doreswamy et al. (2015) employ simulation as a tool to analyse the effects of different RM techniques for different airlines, when switching from leg-based controls to network controls. Cleophas et al. (2009) focus on an approach to

evaluating the quality of RM forecasts both in terms of revenue and common forecast error measurements. Another example of using simulation to evaluate the performance of forecast components is given in Bartke et al. (2018). Temath et al. (2010) used a simulation-based approach to evaluate the robustness of a network-based revenue opportunity model when input data is flawed. In the broader context of demand forecasting, Petropoulos et al. (2014) evaluate fitting time series forecasts for particular patterns of demand evaluation by manipulating these patterns in a simulation framework.

3 Existing Work on Outlier Detection

To assess the existing methodological contributions to outlier detection, we distinguish between identifying outlying observations within a time series (Figure 1a), and identifying an entire outlying time series (in our case, booking pattern) (Figure 1b). In this paper, we aim for the latter.
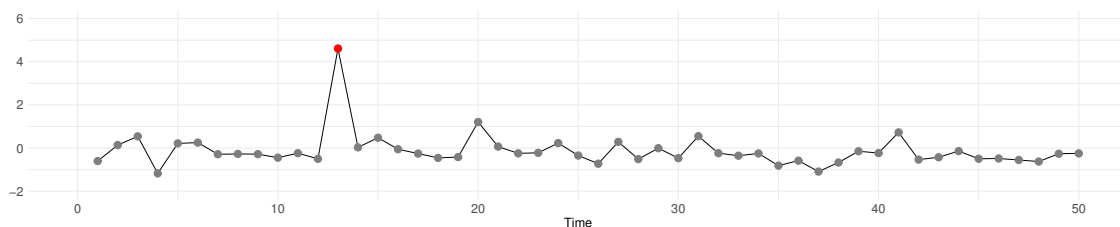
(a) Outlier within a given time series
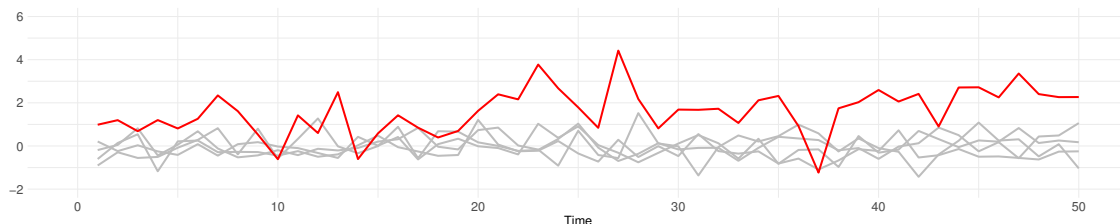
(b) Outlying time series within a collection of series

Figure 1: Different types of outliers in time series data

Literature on handling outliers in the RM process is scarce, though there is some discussion in Weatherford and Kimes (2003): the authors consider removing outliers caused by atypical events, such as holidays and special conventions, to improve future forecasting.

However, they propose only to remove observations outside of the mean $\pm$ 3 standard deviations and do not seek to identify outliers online within the booking horizon.

Beyond RM, a wealth of literature studies outliers (also referred to as anomalies) in time series, as reviewed by Chandola et al. (2009) and Pimentel et al. (2014). For example, Hubert et al. (2015) survey various functional outlier detection techniques for time series data, and apply their methods to multiple real data sets. Barrow and Kourentzes (2018) consider the effect of functional outliers for call centre workload management and recommend an artificial neural network to model them as part of the forecast rather than identifying them. Talagala et al. (2019) propose a sliding window approach for detecting outlying time series within a set of (nonstationary) time series, based on the use of extreme value theory for outlier detection. The authors also distinguish identifying outliers within a time series, and identifying an outlying series from a set. The remainder of this paper distinguishes three classes of approaches to outlier detection: (i) univariate, (ii) multivariate, or (iii) functional. Further technical details of all outlier detection methods described here are available in Appendix A.1.

Univariate Approaches

Univariate outlier detection techniques identify anomalous observations of a single variable, and so can be applied independently at different time points in a time series, e.g., to the cumulative number of bookings per interval in a booking horizon.

- **Nonparametric Percentiles**: This class of approaches uses lower and upper percentiles of the observed empirical distribution at each time point as limits for what constitutes a regular observation as opposed to an outlier. This type of percentile-based approach is discussed by Pincus et al. (1995). It can be used as a basic way to estimate statistics in a more robust manner, by *trimming* or *winsorising* the data (see Dixon and Yuen (1974)). The downside of this approach is that a fixed percentage of the data will always be classified as outliers, even when there are fewer or more actual outliers in the data.

- **Tolerance Intervals**: Statistical tolerance intervals contain at least a specified proportion of observations with a specific confidence level (Hahn and Chandra, 1981). They require two parameters: the coverage proportion, β, and confidence level, $1 - \alpha$. For booking patterns, at each interval of the booking horizon, these approaches define a tolerance interval for the cumulative number of bookings by that time. If the number of

6

observed bookings lies outside of this tolerance interval, the pattern is deemed an outlier. Nonparametric tolerance intervals do not assume an underlying distribution, and instead are based on the order statistics of the data (Wilks, 1941). Parametric tolerance intervals assume an underlying distribution (Hahn and Chandra, 1981). The choice of distribution is not arbitrary, and a bad choice of distribution will perform poorly. Liang and Cao (2018) choose to fit a Normal distribution to hotel booking data to detect anomalous observations.

- **Robust Z-Score**: The Z-score measures where an observation lies in relation to the mean and standard deviation of the overall data (Iglewicz and Hoaglin, 1993). The robust Z-score uses the median and the median absolute deviation to provide a similar measurement. As such, an observation with a robust Z-score above some threshold is classified as an outlier. This score-based method assumes that the observations in a given booking interval are approximately normally distributed based on two justifications: (i) a large proportion of univariate outlier detection methods rely on distributional assumptions (often normality); and (ii) although the discrete, non-negative integer nature of booking data suggests the use of a Poisson distribution, in the presence of trend or seasonal adjustments, the data may no longer have these properties.

Multivariate Approaches

Univariate outlier detection approaches ignore the dependence both within and between time series. We next turn to multivariate approaches as potential methods for capturing within (but not between) time series dependence. In this setting, a time series of length τ, that is, a booking pattern observed over τ intervals, is considered as a point in a τ-dimensional space. This allows the multivariate approaches to compute the distance between any two booking patterns, but ignores the time ordering of observations.

- **Distance**: Each booking pattern (observed over τ intervals) can be characterised by its τ-dimensional distance to every other booking pattern. Aggregating these distances transforms the problem into a univariate outlier detection problem, based on the mean distances. Depending on the length of the booking pattern, issues relating to sparsity due to high dimensionality may arise. As discussed by Aggarwal et al. (2001), some distance metrics perform better than others in a high dimensional space. However, in relation to distance metrics, high dimensionality often refers to at least hundreds of dimensions.

7

The number of booking intervals in RM applications is often fewer than this, ranging from 20 to 50 in examples known to the authors. Therefore, we consider the classical Euclidean and Manhattan distance metrics in our comparative evaluation.

- K-**Means Clustering**: K-means clustering splits the observed booking patterns into K groups by iteratively minimising the (τ-dimensional) distance between each booking pattern and the centre of its assigned cluster (see e.g. MacQueen (1967)). This approach uses a distance threshold to identify booking patterns as outliers based on their distance to the centre of their cluster (Deb and Dey, 2017). As in the distance-based approaches, the choice of distance metric is highly relevant for clustering. Once more, this paper compares Euclidean and Manhattan distance metrics. The approach requires as its input parameter a given K to indicate the number of clusters. Information on the methodology used to determine K is available in Appendix A.1, including a comparison of performance under different choices of K, and the distribution of genuine outliers across such clusters.

Functional approaches

There are two main issues with the use of multivariate outlier detection approaches in this application: (i) the effects of high-dimensionality on distance metrics when considering a large number of booking intervals; and (ii) the lack of accounting for the consecutive, time-ordered, nature of the observations. For such issues, functional data analysis is an intuitive place to turn. Functional data analysis addresses both issues by (i) treating booking patterns as functions observed τ times rather than points in a τ-dimensional space, and (ii) explicitly taking into account the time-ordering of the observations. We provide further analysis of the importance of time-ordering in Appendix C.4.

The functional analysis setting, as discussed by Febrero et al. (2008), lacks a rigorous definition of an outlier. We use the same definition as Febrero et al. (2008): 'a curve is an outlier if it has been generated by a stochastic process with a different distribution than the rest of the curves, which are assumed to be identically distributed'. We view this as a more specific version of the definition by Hawkins (1980) provided in our introduction.

A *depth* function attributes a sensible ordering to observations, such that observations near the centre should have higher depth and those far from the centre should have lower depth. In the functional data setting, this idea provides an ordering to a set of smooth functions observed over discrete time-intervals, with the most central curve trajectory having

8

highest depth. Functional depth not only accounts for the magnitude of the observations, but also for the variability in amplitude i.e. the shape of the curve (Febrero et al., 2008). Given this definition of functional depth, the degree of abnormality of a curve can be characterised by its functional depth, if its depth is particularly low (Hubert et al., 2015). Depth-based approaches for detecting outlying curves are discussed in detail by Hubert et al. (2012). In this paper, we focus on the multivariate halfspace depth described by Claeskens et al. (2014). We state and explain the mathematical definition of the multivariate halfspace depth in Appendix A.1.

4 Proposed Methodology: Functional Outlier Detection With Extrapolation

To improve foresight, RM systems need to identify demand outliers online and as early as possible in the booking horizon. This enables the RM system to update controls for the remainder of the horizon. We term this problem *online outlier detection*. When tasked with online detection at time t_τ in the booking horizon, all approaches discussed in the previous section would exclusively consider the first τ observation intervals only.

Therefore, in the online setting, only a partial booking pattern is available for analysis. We propose to supplement the outlier detection by extrapolating the expected bookings from the current time t_τ up to the end of the booking horizon, t_T. In the computational study, we compare simple exponential smoothing (SES, Chatfield (1975)), autoregressive integrated moving average models (ARIMA, Box and Jenkins (1970)), and integrated generalised autoregressive conditional heteroskedasticity models (IGARCH, Tsay (2002)). Appendix A.2 provides a detailed list of univariate forecasting methods that can be used for extrapolation.

Algorithm 1 outlines the procedure on a set of N booking patterns observed until time t_τ, where given an entire booking horizon of length t_T with $t_1, \ldots, t_\tau, \ldots, t_T$, then $\boldsymbol{y}_n(t_\tau)$ is a time series describing the bookings for pattern n up to time t_τ: $\boldsymbol{y}_n(t_\tau) = \{y_n(t_1), y_n(t_2), \ldots, y_n(t_\tau)\}$.

Figure 2 demonstrates the algorithmic approach; in the extensive simulation analysis, we apply it to a variety of booking patterns and outliers. Figure 2a shows 25 booking patterns that have been observed during the first five of thirty intervals of the booking


**Algorithm 1:** Using extrapolation to improve functional outlier detection
---

1 At time t_τ forecast the accumulation of bookings at each time $t_{\tau+1}, \ldots, t_T$, denoted
  $\hat{y}_n(t_{\tau+1}), \ldots, \hat{y}_n(t_T)$, for each booking pattern n;

2 Calculate $\mathcal{D}_n(\hat{\boldsymbol{y}}_n(t_\tau))$, the functional depth of the observed and extrapolated
  booking pattern $\hat{\boldsymbol{y}}_n(t_\tau) = \{y_n(t_1), y_n(t_2), \ldots, y_n(t_\tau), \hat{y}_n(t_{\tau+1}), \ldots, \hat{y}_n(t_T)\}$, for each
  booking pattern n at time t_τ;

3 Calculate a threshold, C, for the functional depth;

4   Bootstrap the original booking patterns, with probability proportional to
        their functional depths;

5   Smooth the bootstrap samples;

6   Let C^b be the 1st percentile of the depths of the bth bootstrapped sample;

7   Set C as the median value of the C^b;

8 **if** $\mathcal{D}_n(\hat{\boldsymbol{y}}_n(t_\tau)) \leq C$ **then**

9 | Define booking pattern n as an outlier. Delete booking pattern n from the
     | sample of N patterns.

10 **end**

11 **while** $\exists\, n$ *s.t.* $\mathcal{D}_n(\hat{\boldsymbol{y}}_n(t_\tau)) \leq C$ **do**

12 | Recalculate functional depths on the new sample, and remove further outliers.

13 **end**
---

horizon. The extrapolation step is shown in Figure 2b, where the purple lines depict the
ARIMA extrapolation of accumulated bookings until the end of the horizon. The empirical
distribution of the functional depths of the extrapolated sample are shown in Figure 2c,
with the threshold shown in red (computed via the bootstrapping routine described in
Algorithm 1, lines 3-7). The booking patterns classified by the algorithm as outliers are
highlighted in red in Figure 2d.

The input parameters relating to the calculation of the threshold includes the number
of bootstrap samples (line 4), the smoothing method (line 5), and the choice of percentile
(line 6). In this paper, we select parameters as per Febrero et al. (2008) as these perform
well in a wide range of settings. Further details of the threshold calculation are available in
Appendix A.1. The proposed approach could alternatively feature any of the multivariate
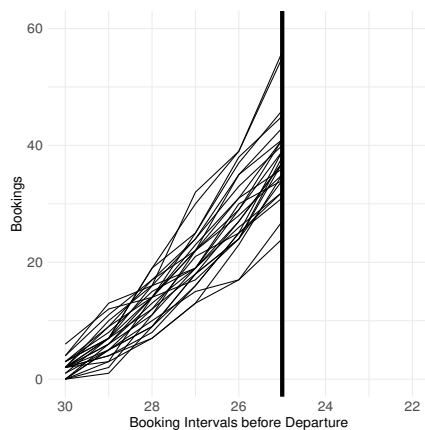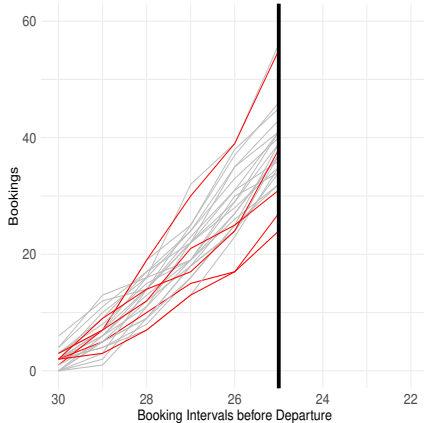
(a) Observed booking patterns up to $t_\tau = 25$
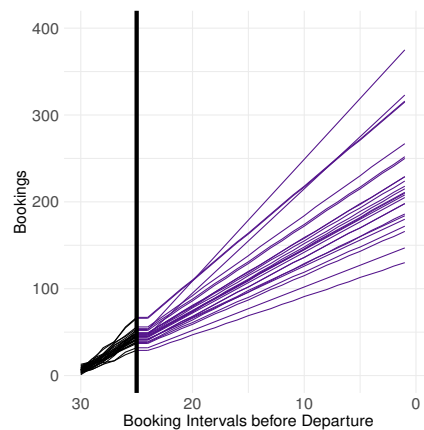
(b) ARIMA extrapolation of booking patterns (purple)
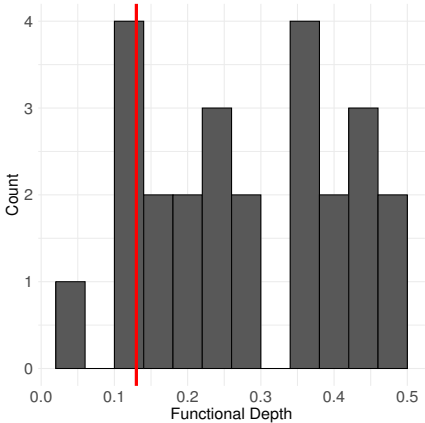
(c) Histogram of functional depths with threshold (red)

(d) Observed booking patterns with detected outliers (red)

Figure 2: Example: functional halfspace depth with ARIMA extrapolation outlier detection

or functional approaches reviewed in Section 3.[1] However, a functional approach provides more scope for extensions, such as considering seasonality and increasing the frequency of outlier detection. In addition, the approach can utilise a variety of methods for extrapolating. Note that the methodology employed for this extrapolation step is independent of the forecasting methodology to predict demand for RM.

---

[1]This approach is not applicable for univariate outlier detection methods as, in this setting, the number of bookings at each point in time is considered independently of past or future bookings.

5 Simulation-based Framework

To quantify effects from demand outliers and evaluate outlier detection approaches, we simulate a basic RM system with capacity controls. Such systems are common in the transport industry, but not limited to that domain (see Talluri and van Ryzin, 2004, Chapter 2.1). The system implemented here is minimal and general and does not fully mirror a real-world application system. However, the booking patterns our simulation generates are comparable with those observed in real-world RM systems – see Appendix C.10. Since the simulation renders the process of demand generation to be explicit, computational experiments can yield truthful detection rates. This is impossible in empirical data analysis, where the true demand and the distinction of *regular* versus outlier demand is never fully certain. Therefore simulation modelling provides an alternative to the problem of creating reproducible forecasting research, highlighted for instance by Boylan et al. (2015).

The simulation implements the following steps:

1. Parameterise a demand model to specify both regular and outlier demand.

2. Generate multiple instances of regular and outlier demand from Step 1 in terms of customer requests (e.g. customers that intend to book a seat on a particular railway connection) arriving across the booking horizon.

3. From the demand model of regular demand (Step 2), compute the forecast in terms of the number of expected requests per fare class and time in the booking horizon.

4. Compute booking limits that maximise expected revenue from bookings based on the demand forecast from Step 3.

5. Use the booking limits (Step 4) to transform arriving requests (from Step 2) into booking patterns over the course of multiple consecutive simulated booking horizons.

6. Analyse booking patterns (from Step 5) to identify booking horizons with outlier demand.

7. Compare knowledge of the underlying demand model (Step 2) to identified outliers (Step 6) to compute detection rates.

	Symbol	Definition	Regular Demand Value
	$\mathcal{I}$	The set of customer types	$\{1 = \text{Business}, 2 = \text{Tourist}\}$
	$\mathcal{J}$	The set of fare classes	$\{\text{A, O, J, P, R, S, M}\}$
	α, β	Parameters of Gamma distribution for number arrivals	$\alpha = 240, \beta = 1$
	a_i, b_i	Parameters of Beta distribution, $\acute{\lambda}_i(t)$	$a_1 = 5, b_1 = 2, a_2 = 2, b_2 = 5$
Fixed	ϕ_i	Proportion of total customer arrivals stemming from type i	$\phi_1 = 0.5, \phi_2 = 0.5$
Input	p_{ij}	Probability of type i being willing-to-pay at most fare class j	$p_{1j} = \{0.35, 0.1, 0.25, 0.15, 0.05, 0, 0\}$ $p_{2j} = \{0.05, 0.1, 0, 0.05, 0.1, 0.15, 0.5\}$
	r_j	Average fare for fare class j	$\{400, 300, 280, 240, 200, 185, 175\}$
	C	Capacity	200
	N_S	Number of runs of simulation used to compute forecasts $\hat{\mu}_j$ and $\hat{\sigma^2}_j$	100
Random	D	Total customer arrivals $\sim Gamma(\alpha, \beta)$	
Input	$\lambda_i(t)$	Time-dependent rate of the Poisson process of type i customer arrivals	
	$x_{n,i,j}(t)$	n^{th} realisation of Poisson process of type i customers purchasing fare class j at time t	
Output	$\hat{\mu}_j$	Forecast of mean of fare class j demand	
	$\hat{\sigma^2}_j$	Forecast of variance of fare class j demand	
	$y_{n,j}(t)$	n^{th} realisation of cumulative bookings in fare class j at time t	

Table 1: Table of notation and parameter values used for simulation

Table 1 sets out the notation used in the remainder of this section to detail the demand model, demand forecasting, revenue maximisation heuristics, and booking limits. In this section, we detail both the models and algorithms, and the parameter settings implemented in the computational study.

5.1 Generating Demand in Terms of Customer Requests

Heterogeneous demand is a frequently stated RM precondition, assuming that customer segments differ in value and can be identified through their idiosyncratic booking behaviour. To model this parsimoniously, the simulation features two customer types but can be easily extended to feature more. We index any parameter that characterises high-value customers with index 1 and any parameter that characterises low-value customers with index 2. Classical RM assumes that requests from high-value customers typically arrive

later in the booking horizon than those from low-value customers. High-value customers are more likely to book expensive fare classes when cheap fare classes are not offered.

We follow Weatherford et al. (1993) in modelling requests from either customer type as arriving according to a non-homogeneous Poisson-Gamma process. Requests from customer type 1 arrive according to a $\text{Poisson}(\lambda_1(t))$ distribution; those from customer type 2 arrive according to a $\text{Poisson}(\lambda_2(t))$ distribution. The total number of customer arrivals D is split between the two segments, such that

$$\lambda_1(t)|(D=d) = d \times \phi_1 \frac{t^{a_1-1}(1-t)^{b_1-1}}{B(a_1,b_1)}, \tag{1}$$

$$\lambda_2(t)|(D=d) = d \times \phi_2 \frac{t^{a_2-1}(1-t)^{b_2-1}}{B(a_2,b_2)}, \tag{2}$$

where $D \sim \text{Gamma}(\alpha, \beta)$ with probability density function:

$$f(d|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)d^{\alpha-1}e^{\beta d}}. \tag{3}$$

The constraint $\phi_1 + \phi_2 = 1$ ensures that all requests belong to exactly one customer type. Additionally, we set parameters a_1, b_1, a_2 and b_2 such that they follow the assumption that valuable customers are more likely to request at later stages of the booking horizon:

$$\frac{a_1 - 1}{a_1 + b_1 - 2} > \frac{a_2 - 1}{a_2 + b_2 - 2}. \tag{4}$$

Figure 3a illustrates arrival rates $\lambda_1(t)$ and $\lambda_2(t)$ across the booking horizon, with Figure 3b showing one realisation of request arrivals in a specific horizon.

A set of fare classes, $1, \ldots, |\mathcal{J}|$, differentiates discount levels, $r_1 \geq r_2 \ldots \geq r_{|\mathcal{J}|}$. The simulation implements a random choice model to let customers choose from the set of currently offered classes. The model assumes all customers book the cheapest available fare class. At the same time, not all customers can afford to book any fare class. For every fare class k, the probability that a customer of type i is willing to pay *at most* fare class k is p_{ik}, as shown in Figure 3d. Each customer has a single fare class threshold, which is the most they are willing to pay, such that:

$$\sum_{k=1}^{|\mathcal{J}|} p_{ik} + p_{i0} = 1, \tag{5}$$

14

where p_{i0} is the the probability of a type i customer arriving and choosing not to book based on the classes on offer. Hence, the probability of booking fare class j is:

$$\mathbb{P}\left(\text{Book fare class } j | \text{No availability in classes } j+1,\ldots,|\mathcal{J}|\right) = \sum_{k=1}^{j} p_k, \qquad (6)$$

$$\mathbb{P}\left(\text{Book fare class } j | \text{Availability in classes } j+1,\ldots,|\mathcal{J}|\right) = 0, \qquad (7)$$

where p_k is the weighted average of probabilities of each customer type i being willing to pay up to fare class k:

$$p_k = \sum_{i \in \mathcal{I}} \phi_i p_{ik}, \qquad (8)$$

and ϕ_i is the proportion of total customer arrivals stemming from type i.

While demand arrival rates vary across the booking horizon, the simulation models arrival rates and choice probabilities as stationary between booking horizons. While, in real-world markets, demand shifts in seasonal patterns and trends, we rely on random draws from distributions with stationary parameters as when introducing and detecting outliers, the simplest case lets all *regular* demand behaviour derive from the same distribution. When an approach cannot correctly detect abnormal demand when all regular demand comes from this same distribution, it is highly unlikely that it will perform better when regular demand is non-stationary.

5.2 Outlier Generation

We generate outlier demand by parameterising demand generation in a way that deviates from the regular setting. Combining outlier demand with booking limits (optimised based on forecasts of regular demand) creates an outlier booking pattern. Outliers can result from three approaches to adjusting the parameters in Equations (1) and (2), and the probabilities p_{ij}:

1. *Demand-volume outliers*: Increasing or decreasing the volume of demand across the whole (or partial) booking horizon, by adjusting the parameters α and β in the Gamma distribution for D, the total demand.

2. *Willingness-to-pay outliers*: Shifting the proportions of demand across fare classes, by either adjusting the choice probabilities per customer type or to the ratio of customer types, ϕ_1, ϕ_2.
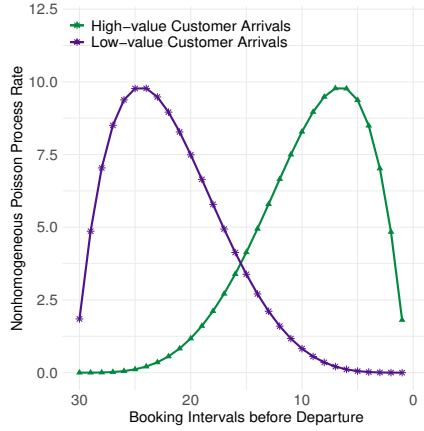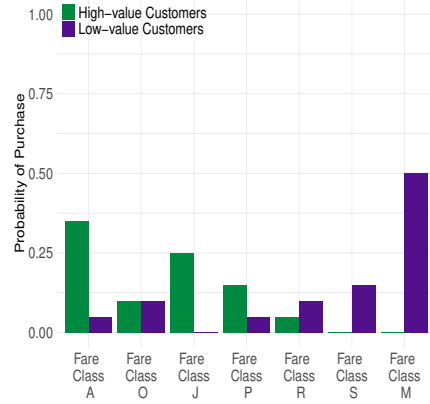
(a) $\lambda_i(t)$, Arrival rates per customer type
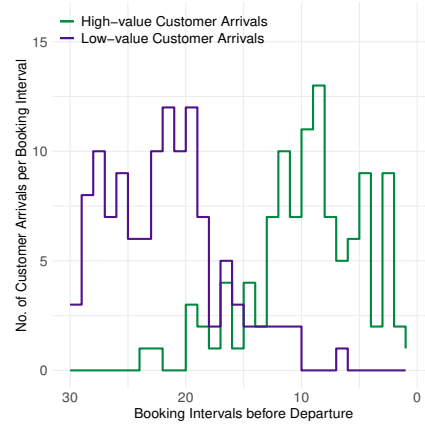
(b) $x_i(t)$, Realisation of the Poisson process with rate $\lambda_i(t)$ per customer type
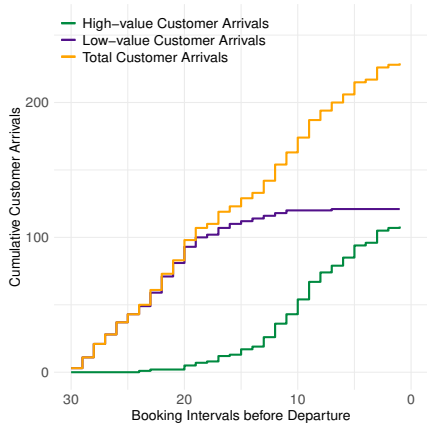
(c) $\sum_{k \leq t} x_i(k)$, Cumulative number of customer arrivals per customer type

(d) p_{ij}, Probability of a customer of type i being willing to pay up to fare class j

Figure 3: Customer arrivals generated by a nonhomogeneous Poisson-Gamma process $D \sim Gamma(240, 1)$, $\phi_1 = \phi_2 = 0.5$, $a_1 = 5$, $b_1 = 2$, $a_2 = 2$, $b_2 = 5$

3. *Arrival-time outliers*: Shifting the arrival pattern of customer requests (from a subset of customer types) over time by adjusting parameters a_1, b_1, a_2, b_2, which control the time at which requests from each customer type arrive.

5.3 Forecasting Demand

Most RM approaches to capacity control rely on knowing the number of expected customer requests per offered product, potentially per set of offered products. The simulation implements heuristics that rely on the mean and the variance of expected requests per fare class (see Section 5.4).

To avoid interference from arbitrary forecasting errors, we exploit knowledge of the demand model given in the simulation setting when creating the forecast. We first draw N_S sets of customer arrivals from Equations (1) and (2). Let $x_{n,i,j}(t)$ define the nth realisation of type i customers who booked in fare class j at time t as drawn from the Poisson arrival process with rate $\lambda_i(t)$ and probability p_{ij}. Then, we set the forecast to be the mean demand across all customer types upon departure from N_S simulations for fare class j, $\hat{\mu}_j$:

$$\hat{\mu}_j = \frac{1}{N_S} \sum_{n=1}^{N_S} \left(\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}} x_{n,i,j}(t) \right). \tag{9}$$

Similarly, the simulation forecasts the variance of the demand for fare class j as:

$$\hat{\sigma_j}^2 = \frac{1}{N_S - 1} \sum_{n=1}^{N_S} \left\{ \left[\left(\sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{I}} x_{n,i,j}(t) \right) - \hat{\mu}_j \right]^2 \right\}. \tag{10}$$

Here we aggregate across the booking horizon in order to obtain forecasts for the final demand for each fare class. The resulting sum of customer requests per fare class across customer types gives the total expected demand per fare class. The mean and variance of these N_S realisations are taken to be the forecasted parameters of a Normal distribution for each fare class demand.

Note that this aggregated forecast deliberately prepares the heuristic applied for revenue optimisation in this case. Applying, for instance, a dynamic program to optimally control arriving customer requests, would require a forecast of customer arrival rates and choice probabilities. The consequence of outliers, however, would be the same, as the arrival rates and choice probabilities deviate for demand outliers.

Last but not least, the simulation forecast assumes stationarity of demand, which is correct with regard to the demand setting simulated here. Therefore, a single forecast value is predicted for all future booking periods. Naturally, in a real-world setting, this stationarity is not given, but instead trends and seasonality complicate forecasting. Future research featuring such forecast aspects would open the path to further differentiation with regard to the effects of different types of outliers given different parameterisations of the non-stationary components.

5.4 Heuristic Revenue Optimisation

The simulation implements two well-known heuristic methods for obtaining booking controls for a single resource: EMSRb and EMSRb-MR. We pick these heuristics for their wide acceptance and pervasive use in practice. Furthermore, as opposed to e.g. exact dynamic programming formulations, these heuristics yield the booking limits widely implemented in current practice. We expect the nature of these booking limits and their updates to be a relevant factor for the recognition and compensation of demand outliers.

- *EMSRb*, Expected Marginal Seat Revenue-b, was introduced by Belobaba (1992). EMSRb calculates joint protection levels for all more expensive classes relative to the next cheaper fare class, based on the mean expected demand and its variance.

- *EMSRb-MR*: To make the EMSRb heuristic applicable when demand depends on the set of offered fare classes, e.g. when customers choose the cheapest available class, Fiig et al. (2010) introduce this variant. It applies a marginal revenue transformation to demand and fares before calculating the EMSRb protection levels based on transformed fares and predicted demand.

Booking limits can be implemented in either a *partitioned* or *nested* way (Brumelle and McGill (1993), and Talluri and Van Ryzin (2004), Chapter 2). Partitioned controls assign capacity such that each unit can only be sold in one specific fare class. Conversely, nested controls let assignments overlap in a hierarchical manner; i.e. units of capacity assigned to one fare class can also be sold in any more expensive fare class. Thus, nested booking limits ensure that for any offered class, all more expensive classes are also offered—as this seems an intuitive goal these booking limits are much more commonly used. Therefore, our simulations implement nested controls.

5.5   Evaluation of Outlier Detection

We regard outlier detection as a binary classification problem, where the two classes are *regular booking patterns* and *outlier booking patterns*. By definition, for any pattern generated in the simulation, we know the true class, as we know the underlying demand model.

Several indicators evaluate the performance of binary classification outcomes, as surveyed by Tharwat (2018). Each outcome falls into one of four categories: (i) if a genuine outlier is correctly classified, it is a *true positive (TP)*; (ii) if a regular observation is correctly classified, it is a *true negative (TN)*; (iii) if a regular observation is wrongly classified as an outlier, it is a *false positive (FP)*; and (iv) if a genuine outlier is wrongly classified as regular, it is a *false negative (FN)*.

To analyse results in this paper, we implement the **Balanced Classification Rate (BCR)** as suggested by Tharwat (2018). This indicator accounts for both the average of the true positive rate and true negative rate:

$$BCR = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right). \tag{11}$$

The notions of high detection rates (fraction of genuine outliers which are correctly detected) and low false positive rates (fraction of regular observations which are incorrectly labelled as outliers) create conflicting objectives. For example, a high true positive rate does not necessarily indicate a high performing algorithm, if it is accompanied by a high false positive rate. Therefore, combining both into a single figure is useful. Nonetheless, additional results on true positive rates, false positive rates, and positive likelihood ratios (Habibzadeh and Habibzadeh, 2019) are included in Appendices C.4 and C.5. Typically, the number of outliers is outweighed by the number of normal observations. This leads to one class being significantly larger than the other. BCR is robust to this imbalance.

Additionally, we generate a **receiver operating characteristic (ROC)** curve by plotting the true positive rate against the false positive rate (McNeil and Hanley, 1984). This provides an additional diagnostic for binary classifiers. The ROC curve compares the true positive to false positive ratio as the threshold (at which an outlier is classified) varies. The optimal ROC curve is that with the combination of highest true positive rate and the the lowest false positive rate, i.e. with an area under the ROC curve closest to 1.

5.6 Experimental Setup

We vary two main elements of the experimental setup for experimental analysis. The first is the parameter settings used to generate regular and outlier demand. The second are the settings of outlier detection.

We generate regular demand according to the parameters in Table 1, which results in regular total demand with a mean of 240, and a standard deviation of 15.492. We benchmark detection performance on outlier demand generated in various ways. Our main focus is on analysing different magnitudes of demand-volume outliers. Our choice of parameter changes for outlier generation follows Weatherford and Pölt (2002), who investigate the effects of inaccurate demand forecasts on revenue. In particular, they consider cases where forecasts are 12.5% and 25% higher or lower than the actual demand. We perform a similar analysis on the benefits of detecting outliers where the overall number of customers deviates from regular demand by $\pm$ 12.5% and $\pm$ 25%. These four types of demand-volume outliers we consider are generated by varying the parameters α and β as described in Table 2. This results in a change in mean of the desired magnitude and direction, but no change in variance. In addition, we consider other types of outliers, as outlined in Section 6.3.

	Mean	Std. Dev	α	β
Regular Demand	240	15.492	240	1
25% Increase	300	15.492	375	1.25
12.5% Increase	270	15.492	303.75	1.125
12.5% Decrease	210	15.492	183.75	0.875
25% Decrease	180	15.492	135	0.75

Table 2: Parameter choices used to generate demand-volume outliers

In a wide-ranging computational study, we compared the performance of all outlier detection methods described in Section 3. Appendix B, Table 3 lists the aggregated results from all experiments carried out. For conciseness, the results discussed in the next section focus on the *best* univariate method, *parametric (Poisson) tolerance intervals*; the best multivariate method, *K-means clustering with Euclidean distance*; the best functional method, *functional depth*; and the best extrapolation method, *ARIMA extrapolation combined with functional depth*.

20

The settings used for these four methods are as follows:

- *Parametric tolerance intervals*: The distribution chosen is Poisson, see Appendix A.1 for details. The coverage proportion is chosen to be $\beta = 0.95$, and the confidence level is $\alpha = 0.05$ by default.
- *K-means clustering*: The number of clusters, K, is chosen to be 2, see Appendix A.1 for reasoning. The default threshold for classifying a booking pattern as an outlier is half the sum of the maximum and minimum distances of the patterns from their cluster centres (Deb and Dey, 2017).
- *Functional depth*: The number of bootstrap samples for the threshold is chosen to be 1000. The smoothing method is as suggested by Febrero et al. (2008). Similarly, the percentile chosen for this analysis is the 1st percentile, as suggested by Febrero et al. (2008).
- *Functional depth + ARIMA extrapolation*: Thresholds are calculated as in functional depth. The orders of the ARIMA extrapolation are selected using `auto.arima` in R, based on AICc, with the augmented Dickey-Fuller test used to choose the order of differencing. The parameters are estimated using maximum likelihood with starting values chosen by conditional-sum-of-squares.

We provide further details on the extent of the computational study, including aggregated results, in Appendix B.

6   Simulation Results

To investigate different outlier simulation and detection techniques, we follow a four-step process. We contrast foresight detection performance of different outlier detection methods in Sections 6.1 and 6.2. This analysis focuses on detection performance across the booking horizon, and evaluates the detection approaches' ability to detect outliers early in the booking horizon. We also quantify the gain in outlier detection performance resulting from the inclusion of the extrapolation step proposed in Section 4. Subsequently, Section 6.3 investigates the effect of different types of outliers on the performance of the outlier detection method. Additionally, Section 6.4 considers an empirical data set to demonstrate the practical implications of the approach. Finally, Section 6.5 presents a final set of experiments intended to measure the potential increase in revenue generated by analysts

correctly taking actions based on alerts from the proposed method of outlier detection. Note that all experiments analysed in this section implement the EMSRb-MR heuristic, which is a better fit with the given demand model. We have investigated the implications of applying the EMSRb heuristic and assessed the revenue generated as well as the effect on identifying outliers in an ancillary study. The results under EMSRb (found in Appendix C.1.) and EMSRb-MR were found to yield similar conclusions, regardless of the outlier detection method used. Additional results, including those relating to the hindsight detection of outliers, are available in Appendix C.

6.1 Benchmarking Foresight Detection of Demand-volume Outliers
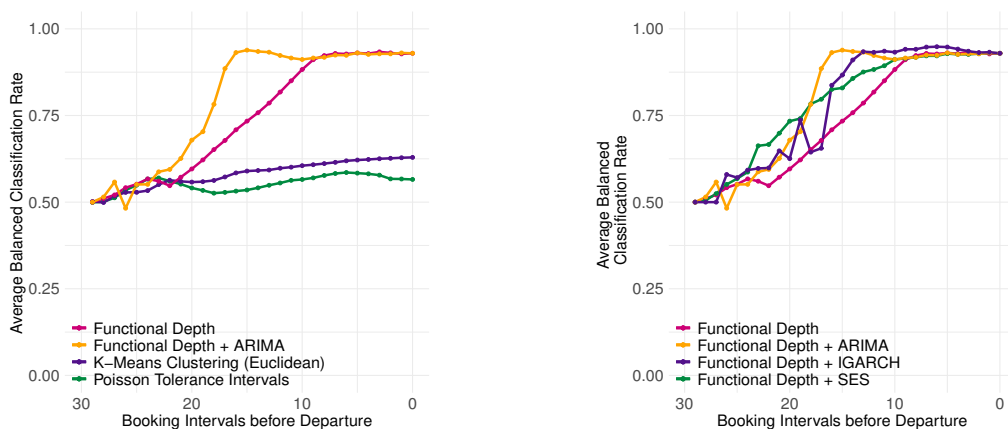
(a) Comparison of best performing outlier detection methods

(b) Improvement from incorporating extrapolation

Figure 4: Comparison of foresight outlier detection averaged over different magnitudes of demand outliers with 5% outlier frequency

To evaluate *foresight detection* performance, Figure 4a displays the average BCR per booking interval. Very early in the booking horizon, all four methods suffer from poor performance but for different reasons – some suffer from low true positive rates, others from high false positive rates (see Appendices C.4 and C.5 for details). At around 21 booking intervals before departure, the average BCR of functional methods quickly accelerate towards 1, whereas the univariate and multivariate approaches at best only show mild improvements in classification performance.

Including ARIMA extrapolation markedly accelerates classification performance, especially between 20 and 10 booking intervals before departure. Note that the aim of extrapolation is not necessarily to increase the overall BCR, but to achieve peak performance earlier in the booking horizon to gain time for analyst intervention. The extrapolation achieves this by increasing the variance of the booking patterns, leading to an increase in the number classified as outliers. See Appendix C.6 for further details. Additional analysis of Receiver Operating Characteristic (ROC) curves (see Section 6.2), further supports the inclusion of ARIMA extrapolation. In Figure 4b, we also compare functional depth with IGARCH and SES extrapolation, and similar improvements are observed as with ARIMA extrapolation. ARIMA provides overall larger gains in performance compared to SES and IGARCH. This is likely due to the flexibility of ARIMA in capturing the changing curvature of the booking pattern, and its ability to encapsulate the autocorrelations induced by censoring from the booking limits. In the last third of the booking horizon, the extrapolation makes up a much smaller part of the pattern, i.e. most of the pattern is now made up of observed rather than extrapolated data. Hence, the input data to the outlier detection algorithm with different extrapolations is similar, and so they produce similar results. Further analysis on the relationship between extrapolation accuracy and the resulting improvement in outlier detection is available in Appendix C.8.

As noted in Section 4, extrapolation could also be combined with multivariate outlier detection methods such as K-means clustering. Given the superior performance of functional depth we focus our main results on this combination, but additional results regarding combining with multivariate techniques are presented in Appendix C.3.

## 6.2   Receiver Operating Characteristic (ROC) Curves

To show that our conclusions in Section 6.1 are robust to different parameterisations of the outlier detection settings, we perform an ROC curve analysis by varying the thresholds for K-means clustering, functional depth, and functional depth with extrapolation. Figure 5 shows the results for two time intervals in the booking horizon: one early at 20 intervals before departure, and one later at 10 intervals before departure.

There are three main conclusions that can be drawn from the results of the ROC analysis. (i) The area under the ROC curve is consistently higher for functional approaches than for K-means. Similarly, the area under the curve is even higher when we include
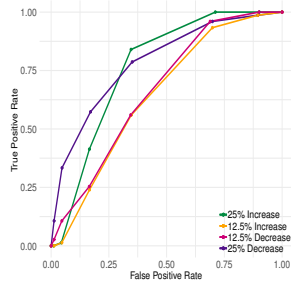
extrapolation. (ii) For K-means, the area under the ROC curve diminishes as the number of booking intervals increases, suggesting issues with sparsity caused by high dimensionality.
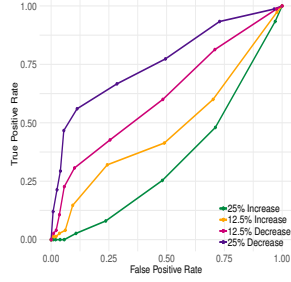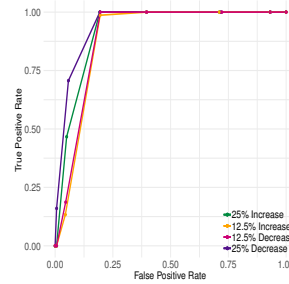

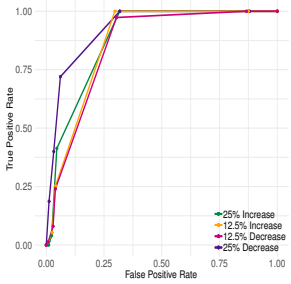(a) K-means clustering outlier detection at 20 booking intervals before departure
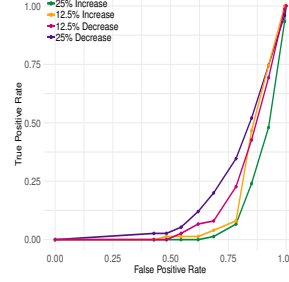
(b) K-means clustering outlier detection at 10 booking intervals before departure
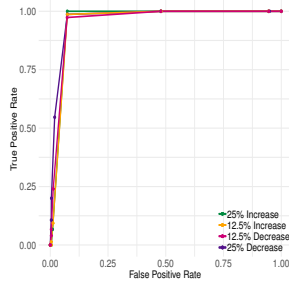
(c) Functional depth outlier detection at 20 booking intervals before departure

(d) Functional depth outlier detection at 10 booking intervals before departure

(e) Functional depth with ARIMA extrapolation outlier detection at 20 booking intervals before departure

(f) Functional depth with ARIMA extrapolation outlier detection at 10 booking intervals before departure

Figure 5: Receiver operating characteristic (ROC) curves

Thus, even a better choice of threshold criteria would not result in improved perfor-

mance for K-means. (iii) The improvement between functional depth and functional depth with extrapolation is smaller towards the end of the booking horizon. This is due to the fact that, at this point, the extrapolation makes up a smaller part of the input data and so the two approaches are more similar.

6.3 Outlier Detection for Diverse Types of Outliers

We next investigate how the average BCR varies depending on the type and magnitude of outliers. All experiments in this section feature an outlier frequency of 5%. When we tested the sensitivity of approaches to different frequencies of outliers (ranging from 1% to 10%, results omitted here), we found little impact on outlier detection performance across methods, such that the conclusions drawn from this section are generally robust. Results on the effect of outlier frequency are available in Appendix C.2.
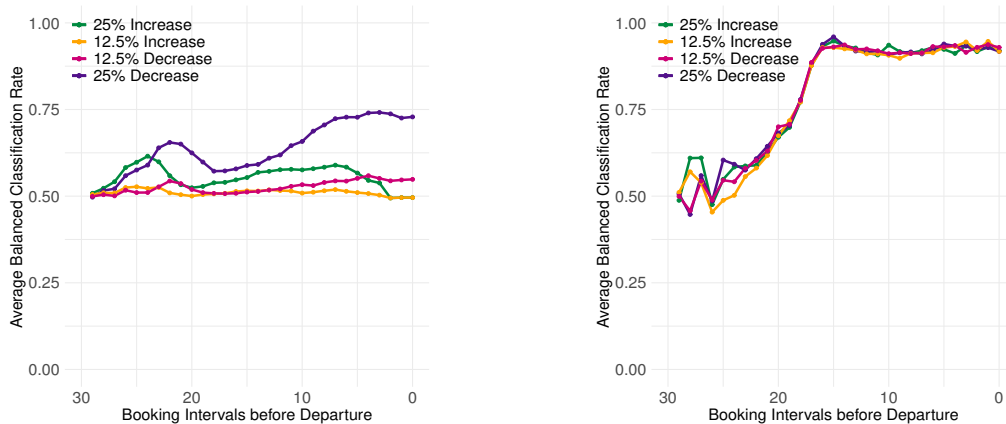
(a) Parametric (Poisson) tolerance intervals outlier detection

(b) Functional halfspace depth with ARIMA extrapolation outlier detection

Figure 6: Balanced Classification Rate under different magnitudes of outliers with 5% outlier frequency

First, we vary the magnitude of *demand-volume outliers* to $\pm 12.5\%$ and $\pm 25\%$. Figure 6a displays the average BCR over time for parametric (Poisson) tolerance intervals. We observe that higher magnitudes of outliers are easier to classify, but also that demand decreases are easier to classify than increases. The latter observation is intrinsic to RM systems: an unexpected decrease in demand causes a decrease in bookings, but an increase

25

in demand does not necessarily result in an increase in bookings if the booking limit for a fare class has been reached, i.e. if the fare class is no longer offered. This censoring leads to the phenomenon of observing a constrained version of demand.

Similar observations arise when testing all other univariate and multivariate outlier detection approaches. In contrast, Figure 6b displays the average BCR over time with functional halfspace depth and ARIMA extrapolation. Here the average BCR is very similar for all four magnitudes of outliers considered. This classification approach therefore appears to be very robust to the magnitude and direction of outliers considered. The robustness to the direction of the outlier demand shift is a result of the choice of depth measure. Hubert et al. (2012) define the multivariate functional halfspace depth for the purposes of identifying curves which are only outlying for a fraction of the time they are observed over. This means that if a booking pattern is affected by censoring, as long as it has still been an outlier before censoring came into effect, it can still be detected later in the horizon. In terms of robustness to magnitude, we hypothesise that much smaller outlier magnitudes would need to be considered before the average BCR decreases. We further consider demand shifts of $\pm 1\%$, $\pm 5\%$, $\pm 10\%$. The results are as expected - for $\pm 10\%$, the performance is only slightly poorer; for $\pm 5\%$, we see a drop in performance with the algorithm at best having a BCR of around 0.75; and a level of $\pm 1\%$ performance is particularly poor with a BCR of close to 0.5. This is behaviour we would expect, given that outliers caused by such a small deviation in demand are unlikely to be considered *outliers* in any real sense. These results are available in Appendix C.7.

Figures 7a and 7b illustrate effects from *willingness-to-pay outliers*, where the ratio of high-value to low-value arrivals changes. The default value in our simulations is $\phi_1 = \phi_2 = 0.5$ such that there is a 1:1 ratio, but we allow this ratio to change to create outliers. Here, $\phi_1 < 0.5$ creates a higher percentage of total arrivals from low paying, early arriving customers of type 2. Under functional depth outlier detection, it is easier to detect this type of outlier when the change in ϕ_1 is larger. There is a dip in performance around interval 22, as this large number of low-paying arrivals causes censoring when booking limits render cheaper classes unavailable. Setting $\phi_1 > 0.5$ creates a larger percentage of type 1 customers, who arrive late and are willing to pay more. Again, this is easier to detect under functional depth when the change in ϕ_1 is larger. Incorporating the ARIMA extrapolation generally improves performance in the last two-thirds of the horizon. However, early in the booking horizon this provides mixed results.
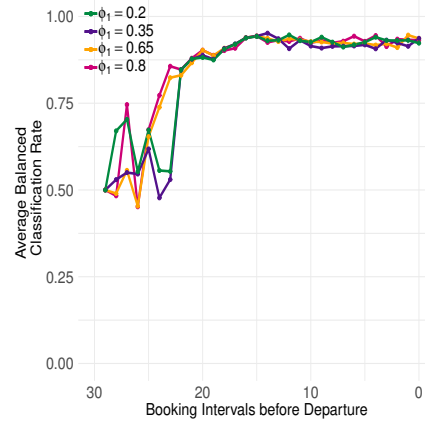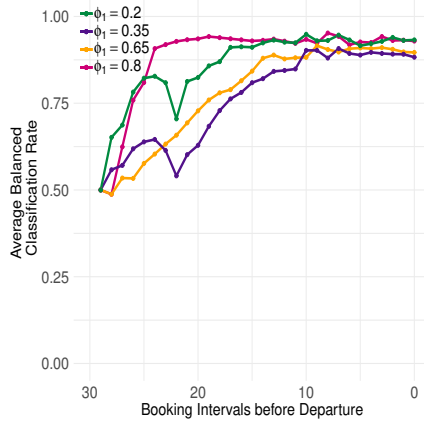
(a) BCR of functional depth outlier detection for changes in ϕ_i

(b) BCR of functional depth with ARIMA extrapolation outlier detection for changes in ϕ_i
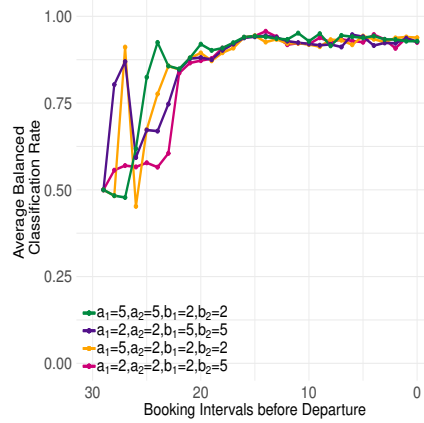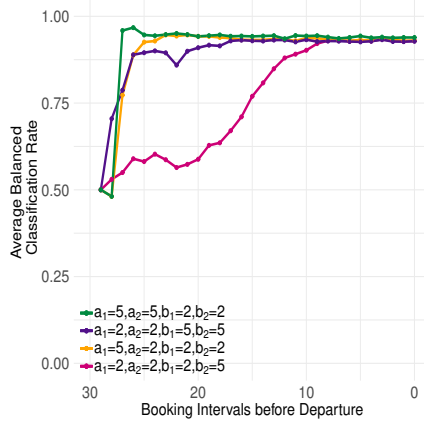
(c) BCR of functional depth outlier detection for changes in a_1, b_1, a_2, b_2

(d) BCR of functional depth with ARIMA extrapolation outlier detection for changes in a_1, b_1, a_2, b_2

Figure 7: Performance of functional depth (with and without ARIMA extrapolation) for different types of outliers

27

Figures 7c and 7d demonstrate the performance of functional depth (with and without extrapolation) for detecting *arrival-time outliers*. These outliers are caused by changes in the parameters a_1, a_2, b_1, b_2 (resulting in a subset of customer types arriving later or earlier than in the regular case), as outlined in Table 3.

	a_1	b_1	a_2	b_2	**Effect of parameter choices**
Regular Demand	5	2	2	5	low value customers arrive before high value customers
Setting 1	5	2	5	2	some low value customers arrive a lot later
Setting 2	2	5	2	5	some high value customers arrive a lot earlier
Setting 3	5	2	2	2	some low value customers arrive a little later
Setting 4	2	2	2	5	some high value customers arrive a little earlier

Table 3: Parameter choices used to generate arrival time outliers

Outliers in Settings 1, 2 and 3 are easy to detect even early in the booking horizon using functional depth without extrapolation. This is fairly intuitive - Settings 1 and 3 create almost no bookings early in the horizon, which is very different from regular behaviour. In contrast, Setting 2 creates far more bookings early in the horizon than the regular setting. ARIMA extrapolation is not needed nor beneficial in Settings 1-3, due to the ease of spotting outliers immediately. In contrast, outliers from Setting 4 are more difficult to detect. This is likely due to the fact that for most of the first half of the horizon, outlier booking patterns and regular booking patterns are similar. In the later half of the horizon, booking limits render the cheaper fare classes unavailable, so that arriving customers purchase higher fare classes only slightly earlier in time. In Setting 4 extrapolation is found to significantly help the classification performance in this more challenging setting.

6.4   Detecting Outliers in Railway Booking Patterns

We demonstrate the proposed outlier detection method by identifying outliers in a data set of 1387 booking patterns obtained from the main German railway company, Deutsche Bahn. This preliminary empirical study can be thought of as a guide to practitioners for how to apply the algorithm. A detailed analysis of the algorithm's performance in practice would require a manually annotated data set or, potentially, a field study. While of significant interest, such an analysis is beyond the scope of this paper.

We consider booking patterns that were observed for a single departure time every day of the week, for one railway leg that directly connects an origin and a destination. The 1387 booking patterns are observed over 18 booking intervals, where the first booking interval is observed 91 days before departure. Figure 8a illustrates 148 of these booking patterns, which relate to trains departing on Mondays. For the purposes of Figure 8a, we have rescaled the number of bookings to be between 0 and 1. The booking data is generated from an RM system that implements an EMSR variant, which sets and updates booking limits based on forecasted demand and observed bookings.

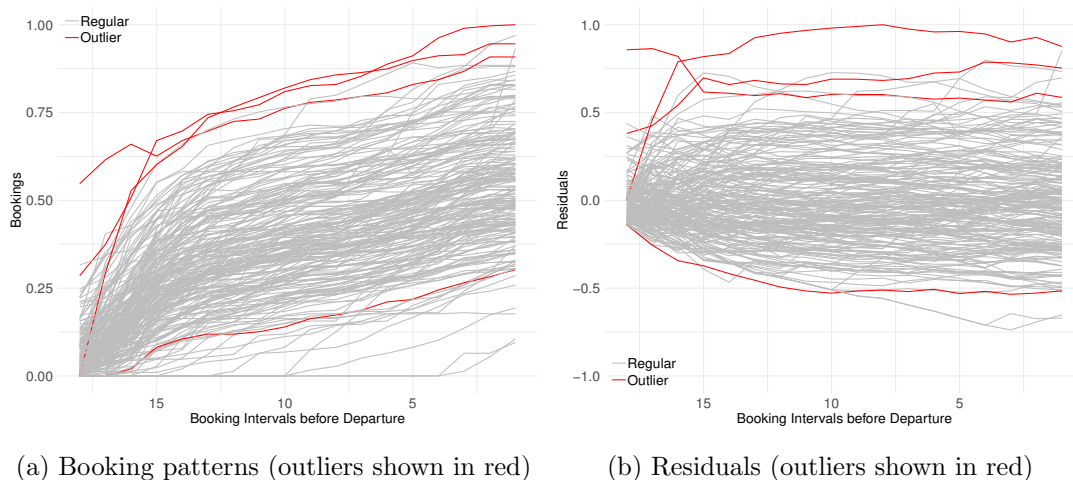(a) Booking patterns (outliers shown in red) (b) Residuals (outliers shown in red)

Figure 8: Pre-processing of data

In order to obtain a homogeneous data set to allow for outlier detection, we must account for two factors: (i) departure days of the week and (ii) shortened booking horizons. We compare booking patterns for different days of the week by applying pairwise functional ANOVA tests (Cuevas et al., 2004). In general, booking patterns for different days of the week are not directly comparable (see Appendix C.10 for details). In addition, shortened booking horizons are a special characteristic of this data set that are caused by the railway service provider's process for implementing schedule changes. As a consequence, some booking horizons are foreshortened and the majority of bookings typically arrive much closer to departure (see Appendix C.10).

To prepare the data for outlier detection, we transform the booking patterns to make them more comparable to each other. To account for both shortened booking horizons and

departure days of the week, we apply a functional regression model (Ramsay et al., 2009). This functional regression model accounts for the way in which average booking patterns change from day to day, and fits a mean function (see Appendix C.10 for details) to the booking patterns for each day of the week. The model is of the form:

$$bookings_i(t) = \beta_0(t) + \beta_1(t)I_{Monday_i} + \beta_2(t)I_{Tuesday_i} + \beta_3(t)I_{Wednesday_i} +$$
$$\beta_4(t)I_{Thursday_i} + \beta_5(t)I_{Friday_i} + \beta_6(t)I_{Saturday_i} + \beta_7(t)I_{Shorter\ Horizon_i} + e_i(t), \tag{12}$$

where the $\beta_j(t)$ are functions of time. Here, $I_{Monday_i} = 1$ if booking pattern i relates to a departure on a Monday, 0 otherwise, and so on. Since every departure belongs to a single day of the week, $\beta_0(t)$ represents the average bookings for Sunday departures, with a non-shortened booking horizon. This means that $\beta_1(t)$ accounts for the change in average bookings between Sunday and Monday departures. The purpose of allowing the $\beta_j(t)$ to be functions of time is not to remove the trend from the booking patterns but rather to allow the relationship between different days of the week to change over the course of the booking horizon. In this model, $I_{Shorter\ Horizon_i} = 1$ if the booking horizon has been shortened due to scheduling changes (affecting departures from mid-December to mid-March), 0 otherwise.

We run the functional depth outlier detection routine on the residuals, as shown in Figure 8b, with detected outliers shown in red. We also show these corresponding outliers in red in Figure 8a. Of the 1387 booking patterns in the data set, we classify 66 ($\approx 5\%$) as outliers. Note that the frequency of outliers is not an assumption provided to the outlier detection routine, and coincides with the frequency of outliers used in the simulation setup (5%), thus justifying this choice in our earlier simulations.

For validation, we provided the labelled data set back to Deutsche Bahn. The company's domain experts have confirmed that the relative proportion of outliers is appropriate to support analyst work on improving demand forecast and booking controls. Furthermore, their hindsight analysis has confirmed that most automatically identified outliers would have benefitted from such corrections.

In addition, we compared the dates of the booking patterns classified as outliers with a list of known holidays and events. Of the 66 booking patterns classified as outliers, 30 could be attributed to known events e.g. public holidays. This leaves 36 outlying booking patterns which would otherwise have gone undetected. However, we do not aim to solely identify already known events, as there would be little point to only confirming

30

known information. Therefore, the additionally identified outliers are not necessarily false positives – they are in fact the very booking patterns we are attempting to identify.

6.5 Revenue Improvement Under Outlier Detection of Demand-volume Outliers

To evaluate the effect of demand deviating from the forecasts used by EMSRb and EMSRb-MR, we now introduce a best-case scenario where the RM system anticipates outliers and generates accurate demand forecasts (as opposed to implementing booking controls based on the initial erroneous forecasts). The percentage change in revenue, when switching from erroneous to correct forecasts, under four demand changes is shown in Table 4. Results show the impact of detecting and correcting outliers in demand depends on the demand factor, the choice of booking control heuristic, and the magnitude of the demand deviation.

Under EMSRb, the effect on revenue is asymmetric across positive and negative outliers. When the outlier is caused by a decrease in demand, correcting the forecast and updating controls leads to significant increases in revenue, particularly at higher demand factors. Conversely, when the outlier is caused by an increase in demand, correcting the forecast and updating controls has a negative impact on revenue. Although counter-intuitive at first glance, this agrees with previous findings. EMSRb is known to be too conservative (Weatherford and Belobaba, 2002) and reserve too many units of capacity for high fare classes, thereby rejecting an excessive number of requests from customers with a lower willingness to pay. In consequence, there is left-over capacity at the end of the booking horizon. Hence, under-forecasting can be beneficial under EMSRb.

Under EMSRb-MR booking controls, the results are more symmetric across positive and negative outliers, in that correctly adjusting forecasts increases revenue regardless of whether the initial forecast was too high or too low. Under both types of heuristic, the magnitude of the change in revenue (either positive or negative) is generally larger when the change in demand from the forecast is larger.

Figure 9 shows the average percentage gain in revenue, at each point in the booking horizon, from analysts correcting forecasts for those booking patterns identified as outliers. The percentage gain is in comparison to the analyst making no changes and using the incorrect forecast for the entirety of the booking horizon.

The outlier detection method of choice in Figure 9 is functional depth with ARIMA

Optimisation Heuristic	Forecasted Demand Factor	% Change in Demand from Forecast			
		-25%	**-12.5%**	**+12.5%**	**+25%**
EMSRb	0.90	+0.1%	+0.1%	-0.9%	-3.6%
	1.20	+10.2%	+6.4%	-2.3%	-2.3%
	1.50	+12.2%	+4.4%	-4.5%	-6.8%
	Avg.	**+7.5%**	**+3.6%**	**-2.5%**	**-4.2%**
EMSRb-MR	0.90	+2.3%	+1.3%	+0.4%	+2.9%
	1.20	+2.0%	+4.1%	+4.4%	+9.9%
	1.50	+16.2%	+7.7%	+5.0%	+9.5%
	Avg.	**+6.9%**	**+4.4%**	**+3.3%**	**+7.4%**

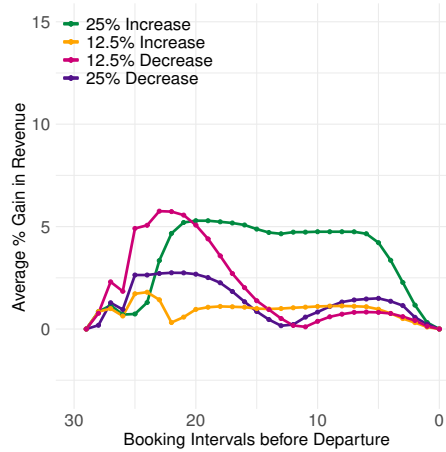Table 4: % Change in revenue resulting from correcting inaccurate demand forecasts

Figure 9: Gain in revenue under different magnitudes of outliers using functional depth with ARIMA extrapolation

extrapolation. We consider an idealised scenario, in that when a booking pattern is flagged as an outlier, if it is a true positive (genuine outlier) then analysts adjust the forecast according to the correct distribution. Similarly, if the flagged outlier is a false positive, analysts do not make any changes to the forecast. Although idealised, the results here highlight the potential gains in revenue from analyst intervention, as well as the utility of using functional outlier detection in detecting true positives and avoiding false negatives (missed outliers).

Results show the use of our method creates a peak early in the booking horizon, when the potential revenue gain is highest. This peak is caused by a combination of being far enough into the booking horizon such that some bookings have occurred and the outlier detection method is able to identify outliers, but being early enough in the horizon such that any actions taken still have time to make an impact.

7   Conclusion and Outlook

In conclusion, the work presented in this paper gives rise to several insights.

We benchmarked a set of outlier detection techniques and find that the functional outlier detection approach offers the best performance and the most scope for further extensions. Our results show that combining functional outlier detection with our proposed extrapolation step significantly improves performance overall, and accelerates the correct identification of outliers earlier in the booking horizon. We do note however that all methods perform poorly very early in the booking horizon where very little data has been gathered, and clearly at this stage analyst expertise or prior information is needed rather than relying on booking data alone.

By analysing an empirical railway booking data set, we demonstrated that such data is similar in shape as the data generated by the simulation model. Furthermore, the frequency of outliers detected via applying functional outlier detection to the empirical data was similar to what was observed with simulation data. In contrast to the simulation setting, the empirical data does not provide information on the labelling of actual outliers, it was therefore not possible to compute detection rates for this data. However, we validated our findings by presenting them to domain experts.

Outliers in demand diminish revenue when they go undetected. The exact effect depends on the combination of outlier and optimisation method, as shown in Section 6.5.

Nevertheless, we argue that using a heuristic with an intrinsic bias that is then compensated by undetected outliers (as observed for EMSRb and undetected positive demand outliers) cannot be desirable for an automated system.

We have demonstrated that identifying outlier booking curves and adjusting the demand forecast accurately early in the booking horizon supports revenue optimisation. Currently, revenue management analysts decide on which booking patterns are outliers based on their previous experience of observing demand and their knowledge about special events. Automated outlier detection routines provide another procedure of alerting analysts to unusual patterns. If the detection algorithm identifies a booking pattern as an outlier, the RM system alerts the responsible analyst. When the system and the analyst agree that a booking pattern is critical and that it requires intervention, an analyst must decide which action(s) to take. Specifically, they need to decide whether to increase or decrease the forecast or inventory controls, and by how much. Further work could investigate methods to adjust the initial forecast to account for outliers.

Within the context of RM, thoroughly examining the effects across further outlier situations, e.g. outliers affecting only part of the booking pattern, and optimisation solutions, e.g. dynamic programming, is an interesting area for further research. Furthermore, future research might consider more differentiated forecasting situations, featuring trends and seasonalities. Beyond RM, other paradigms of offer optimisation, such as mark-down pricing or the pricing of Veblen products, might offer different challenges with regards to outlier detection. Given that the resulting sales observations should also take the format of time series, we consider it interesting to find out whether the same methods would broadly apply in such different settings.

Acknowledgements

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. *In: Van den Bussche J., Vianu V. (eds) Database Theory. ICDT 2001. ICDT 2001. Lecture Notes in Computer Science, vol 1973. Springer, Berlin, Heidelberg*, pages 420–434.

Banerjee, N., Morton, A., and Akartunalı, K. (2019). Passenger demand forecasting in scheduled transportation. *European Journal of Operational Research*. in press.

Barrow, D. and Kourentzes, N. (2018). The impact of special days in call arrivals forecasting: A neural network approach to modelling special days. *European Journal of Operational Research*, 264(3):967–977.

Bartke, P., Kliewer, N., and Cleophas, C. (2018). Benchmarking filter-based demand estimates for airline revenue management. *EURO Journal on Transportation and Logistics*, 7(1):57–88.

Belobaba, P. P. (1989). OR Practice: Application of a Probabilistic Decision Model to Airline Seat Inventory Control. *Operations Research*, 37(2).

Belobaba, P. P. (1992). Optimal vs. heuristic methods for nested seat allocation. In *Proceedings of AGIFORS Reservations and Yield Management Study Group (1992)*, pages 28–53.

Box, G. and Jenkins, G. (1970). *Time Aeries Analysis: Forecasting and Control.* San Francisco: Holden-Days.

Boylan, J., Goodwin, P., Mohammadipour, M., and Syntetos, A. (2015). Reproducibility in forecasting research. *International Journal of Forecasting*, 31(1):79–90.

Brumelle, S. L. and McGill, J. I. (1993). Airline seat allocation with multiple nested fare classes. *Operations Research*, 41(1):127–137.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Survey of Anomaly Detection. *ACM Computing Survey*, 41(3):1–72.

Chatfield, C. (1975). *The Analysis of Time Series: An Introduction.* Chapman and Hall.

Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014). Multivariate functional half-space depth. *Journal of the American Statistical Association*, 109(505):411–423.

Cleophas, C., Frank, M., and Kliewer, N. (2009). Simulation-based key performance indicators for evaluating the quality of airline demand forecasting. *Journal of Revenue and Pricing Management*, 4(8):330–342.

Cleophas, C., Kadatz, D., and Vock, S. (2017). A Literature Survey of Recent Theoretical Advances. *Journal of Revenue and Pricing Management*, 16(5):483–498.

Cuevas, A., Febrero, M., and Fraiman, R. (2004). An anova test for functional data. *Computational Statistics & Data Analysis*, 47:111–122.

Deb, A. B. and Dey, L. (2017). Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering. *World Journal of Computer Application and Technology*, 5(2):24–29.

Dixon, W. and Yuen, K. K. (1974). Trimming and winsorization: A review. *Statistische Hefte*, 15(2-3):157–170.

Doreswamy, G. R., Kothari, A. S., and Tirumalachetty, S. (2015). Simulating the flavors of revenue management for airlines. *Journal of Revenue and Pricing Management*, 6(14):421–432.

Febrero, M., Galeano, P., and González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics*, 19(4):331–345.

Fiig, T., Isler, K., Hopperstad, C., and Belobaba, P. (2010). Optimization of mixed fare structures: Theory and applications. *Journal of Revenue & Pricing Management*, 9(12):152–170.

Frank, M., Friedemann, M., and Schröder, A. (2008). Principles for simulations in revenue management. *Journal of Revenue and Pricing Management*, 1(7):215–236.

Gönsch, J. (2017). A survey on risk-averse and robust revenue management. *European Journal of Operational Research*, 263(2):337–348.

Habibzadeh, F. and Habibzadeh, P. (2019). The likelihood ratio and its graphical representation. *Biochemia medica*, 29(2).

Hahn, G. J. and Chandra, R. (1981). Tolerance Intervals for Poisson and Binomial Variables. *Journal of Quality Technology*, 13(2):100–110.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall.

Hubert, M., Claeskens, G., De Ketelaere, B., and Vakili, K. (2012). A new depth-based approach for detecting outlying curves. In Colubi, A., Fokianos, K., Gonzalez-Rodriguez, G., and Kontoghiorghes, E., editors, *Proceedings of COMPSTAT 2012*, pages 329–340.

Hubert, M., Rousseeuw, P. J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods and Applications*, 24(2):177–202.

Iglewicz, B. and Hoaglin, D. (1993). The ASQC Basic References in Quality Control: Statistical Techniques. *In: Mykytka, E.F., (eds), How to Detect and Handle Outliers*, 16.

Kimms, A. and Müller-Bungart, M. (2007). Simulation of stochastic demand data streams for network revenue management problems. *OR Spectrum*, 1(29):5–20.

Lawrence, M., Goodwin, P., O'Connor, M., and Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of forecasting*, 22(3):493–518.

Lawrence, M., O'Connor, M., and Edmundson, B. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research*, 122(1):151–160.

Liang, T. X. and Cao, C. X. (2018). Outliers detect methods for time series data. *Journal of Discrete Mathematical Sciences and Cryptography*, 21(4):927–936.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, volume 1, pages 281–297. University of California Press.

McNeil, B. J. and Hanley, J. A. (1984). Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves. *Medical Decision Making*, 4(2):137–150.

Morales, D. R. and Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research*, 202(2010):554–562.

Mukhopadhyay, S., Samaddar, S., and Colville, G. (2007). Improving revenue management decision making for airlines by evaluating analyst-adjusted passenger demand forecasts. *Decision Sciences*, 38(2):309–327.

O'Connor, M., Remus, W., and Griggs, K. (1993). Judgemental forecasting in times of change. *International Journal of Forecasting*, 9(2):163–172.

Pereira, L. N. (2016). An introduction to helpful forecasting methods for hotel revenue management. *International Journal of Hospitality Management*, 58:13–23.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., and Nikolopoulos, K. (2014). 'horses for courses' in demand forecasting. *European Journal of Operational Research*, 237(1):152–163.

Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.

Pincus, R., Barnett, V., and Lewis, T. (1995). Outliers in Statistical Data. 3rd Edition. *Biometrical Journal*, 37(2):256.

Ramsay, J. O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis in R and Matlab*. Springer, New York.

Talagala, P. D., Hyndman, R. J., Smith-Miles, K., Kandanaarachchi, S., and Muñoz, M. A. (2019). Anomaly Detection in Streaming Nonstationary Temporal Data. *Journal of Computational and Graphical Statistics*.

Talluri, K. and van Ryzin, G. (2004). Revenue Management Under a General Discrete Choice Model of Consumer Behavior. *Management Science*, 50(1):15–33.

Talluri, K. T. and Van Ryzin, G. J. (2004). *The Theory and Practice of Revenue Management*. Kluwer Academic Publishers.

Temath, C., Pölt, S., and Suhl, L. (2010). On the robustness of the network-based revenue opportunity model. *Journal of Revenue and Pricing Management*, 4(9):341–355.

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*, pages 1–13. in press.

Tsay, R. (2002). *Analysis of Financial Time Series*. John Wiley and Sons.

Weatherford, L. (2016). The history of forecasting models in revenue management. *Journal of Revenue and Pricing Management*, 15(3-4):212–221.

Weatherford, L. R. and Belobaba, P. P. (2002). Revenue impacts of fare input and demand forecast accuracy in airline yield management. *The Journal of the Operational Research Society*, 53(8):811–821.

Weatherford, L. R., Bodily, S. E., and Pfeifer, P. E. (1993). Modeling the Customer Arrival Process and Comparing Decision Rules in Perishable Asset Revenue Management Situations. *Transportation Science*, 27(3):239–251.

Weatherford, L. R. and Kimes, S. E. (2003). A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting*, 19(3):401–415.

Weatherford, L. R. and Pölt, S. (2002). Better unconstraining of airline demand data in revenue management systems for improved forecast accuracy and greater revenues. *Journal of Revenue and Pricing Management*, 1(3):234–254.

Wilks, S. S. (1941). Determination of Sample Sizes for Setting Tolerance Limits. *The Annals of Mathematical Statistics*, 12(1):91–96.

Zeni, R. H. (2003). The value of analyst interaction with revenue management systems. *Journal of Revenue and Pricing Management*, 2(1):37–46.