Unconstraining Methods for Revenue Management Systems under Small Demand

Nikolaos Kourentzes[a], Dong Li *,[b], and Arne K. Strauss[c]

[a] *Lancaster University Management School, University of Lancaster, Lancaster LA1 4YX, UK*

[b] *The York Management School, University of York, York YO10 5GD, UK*

[c] *Warwick Business School, University of Warwick, Coventry CV4 7AL, UK*

Dr Nikolaos Kourentzes is an Associate Professor in the Department of Management Science, Lancaster University Management School. Nikos researches in several areas of business forecasting and his work has been presented in numerous international academic and practitioner conferences. He is regularly giving talks on improving and automating forecasting in organisations using established and state-of-the-art statistical methods. He has substantial experience in applied research projects with various organisations.

Dr Dong Li is a Lecturer (Assistant Professor) of Operations Management at the University of York. He has rich knowledge and experience in revenue management, especially in travel and hospitality industries. Before the current post he had spent a few years working for the revenue management team in a major rental car company in the UK.

Dr Arne K. Strauss is an Associate Professor of Operational Research at the University of Warwick. He has conducted research on revenue management over the past decade, including various collaborations with industry.

[*] Corresponding author, E-mail address: dong.li@york.ac.uk, Tel: +44 1904 325046 (Dong Li)

Unconstraining Methods for Revenue Management Systems under Small Demand

Abstract

Sales data often only represents a part of the demand for a service product owing to constraints such as capacity or booking limits. Unconstraining methods are concerned with estimating the true demand from such constrained sales data. This paper addresses the frequently encountered situation of observing only a few sales events at the individual product level and proposes variants of small demand forecasting methods to be used for unconstraining. The usual procedure is to aggregate data; however, in that case we lose information on when restrictions were imposed or lifted within a given booking profile. Our proposed methods exploit this information and are able to approximate convex, concave or homogeneous booking curves. Furthermore, they are numerically robust due to our proposed group-based parameter optimization. Empirical results on accuracy and revenue performance based on data from a major car rental company indicate revenue improvements over a best practice benchmark by statistically significant 0.5%–1.4% in typical scenarios.

Keywords: demand unconstraining; forecasting; small demand; revenue management

1 Introduction

Revenue management (RM) systems manage demand for services over an advance booking period by controlling the availability of certain offerings at an operational level, e.g. by using booking limits. They can be found in many industries such as airlines, hotels, car rentals, trains, cruise shipping and many more. Imposing these controls however means that no more sales can be observed once an offering has become unavailable, so the sales volume may significantly underestimate the actual market demand. A significant proportion of booking data is constrained in that way: booking restrictions were in place for 30 per cent of the time (on average) in car rental data available to us. Due to the seminal work of Cooper et al. (2006) it is well-known that using such constrained sales data as input to forecasting can lead to the so-called spiral-down effect: a negative self-reinforcing cycle resulting in decreasing revenue performance of the RM system.

Many methods have been proposed to unconstrain sales data (see Guo et al., 2012, for a review). The need for unconstraining arises in RM regardless of whether only a single resource (e.g., a flight leg) or a whole network is being managed. However, when managing revenue on a network of resources, small sales data is more commonly encountered at the level that constraints are being implemented. In this case, the widely used Expectation Maximization (EM) approach has been shown to perform poorly (Queenan et al., 2007). We likewise focus on this case, and therefore present our approach in the context of network RM, even though the proposed methods work in either situation.

To explain why small data more frequently arises in network RM, we first need to clarify the terminology of resources and products in network RM. For example, a resource in car rental RM corresponds to the inventory of a specific car type available on a specific day at a specific station. The same logic applies to hotels, trains, airlines, equipment hire and other industry sectors using network RM. The term "network" RM stems from having to manage resources simultaneously since products are defined on a network of resources. A product in the example of car rental could refer to the combination of a pick-up date, length-of-rental (LoR), station, car type and booking channel. Each product defined in that way uses resources corresponding to the days that the car is rented out. Typically, network RM applications feature a large number of products defined in this way, especially when the number of resources is large. It is common that many of these products have only a few (less than 10) sales events recorded over the entire booking horizon, which makes the small sales case important for practice. We cannot aggregate these sales figures without losing information on when restrictions came into place; information that can be exploited to improve unconstraining.

We draw on the substantial body of literature on small and intermittent demand forecasting to investigate whether more recent developments on algorithms and parameter optimization are able to improve over the exponential smoothing approaches used by Queenan et al. (2007) for small demand. We propose three key improvements: (i) a generalization of trended exponential smoothing that captures better non-homogeneous booking curves, regardless of whether they are concave, convex or linear; (ii) a novel way to optimize model parameters that improves the performance of both proposed and benchmark methods; and (iii) optimize the smoothing parameters over groups of constrained booking curves so as to improve robustness without losing information on the individual restriction start times, particularly when the sample is very limited.

Their performance is empirically tested against best practice benchmarks. We conduct a simulation study based on actual sales data from a major UK-based car rental company. To gauge revenue impact, we build a slightly simplified version of their in-house developed RM system. We find statistically significant revenue improvements over the best practice benchmarks (Croston and Holt's methods as proposed by Queenan et al., 2007) by 0.5%–1.4%. Revenue improvements on this order are significant because for many relevant industry sectors they translate directly into additional profits due to low marginal costs. The ease of implementation of the proposed methods further adds to their appeal.

The paper is structured as follows: In §2 we review the literature, in §3 we propose and discuss the unconstraining methods, in §4 we present our numerical results regarding accuracy and revenue performance of the proposed methods, and we draw conclusions in §5.

2 Literature Review

Azadeh et al. (2014) present a taxonomy of demand unconstraining methods with four categories, namely basic, statistical, choice and optimization based. Our proposed methods use exponential smoothing and require parameter optimization and hence belong to the last category, along with expectation maximization, projection detruncation (PD) and non-linear programming. They describe expectation maximization as one of the most popular unconstraining methods, however Queenan et al. (2007) report that EM did not perform well on small sample sizes, which we focus on here.

In an earlier review of unconstraining methods, Guo et al. (2012) distinguish essentially three categories, namely single- and multi-product methods for a single resource, and multi-resource / multi-product methods. Our proposed models form part of the single resource, single product literature which represents the majority of papers. The work of Weatherford and Pölt (2002), which belongs to the same category, identifies PD, EM and an averaging method as the best techniques. We use the latter method as a benchmark and refer to it as *Averaging*; details on that method are given in the online supplement. We do not consider PD and EM due to the aforementioned issues on small data samples.

Our work has been motivated by Queenan et al. (2007) who propose to use Holt's exponential smoothing method (named double-exponential smoothing therein) and Croston's method for unconstraining. They show that these methods perform well against PD, EM and *Averaging.* Although EM is a strong contender on larger data sets, they find that exponential smoothing is better when little historical data is available or all demand sets are constrained. They use the observed sales over unconstrained time periods to estimate smoothing parameters by minimizing the sum of the squared errors in-sample for each booking curve individually. Demand over constrained time periods is estimated by extrapolating linearly using the most recent trend estimate (Holt) or demand arrival rate (Croston). That way, their approach takes into account when restrictions were imposed (or lifted) as opposed to other methods that ignore this time aspect. We use their proposed Holt's and Croston's methods as benchmarks.

For non-homogeneous booking curves (e.g. concave ones as encountered in resorts where customers book long in advance, or convex ones as in urban car rental stations where people book shortly prior to pick up), the use of a linear forecast is counter-intuitive. Therefore, we suggest a modification of the damped trend exponential smoothing method (Gardner and McKenzie, 1985), since this allows one common framework for all shapes of booking curves (convex, concave or homogeneous). For an overview of exponential smoothing methods, see Gardner (2006).

As we are especially interested in unconstraining small data sets, we consult the literature on small and intermittent demand forecasting for further improvements. Exponential smoothing methods and variants are widely used due to their simplicity and effectiveness in the face of small data samples (for examples, see: Willemain et al., 1994; Babai et al., 2012; Bacchetti and Saccani, 2012). Croston's method, which itself is based on exponential smoothing, is widely regarded as more appropriate for this type of data (Syntetos and Boylan, 2005). Croston's method is biased, as discussed by Syntetos and Boylan (2001), and the Syntetos Boylan Approximation (SBA) by Syntetos and Boylan (2005) addresses this issue. However Kourentzes (2014) provides evidence that any difference between the original Croston's and SBA are minimal when appropriately optimized, as the bias is very strong only for relatively large parameters. Teunter et al. (2011) criticize Croston's method for updating its estimates only after a positive demand event, ignoring the in-between periods with clear implications for obsolescence. To address that they model the demand probability instead of the demand interval, as in Croston's method, which can be updated every period, resulting in the Teunter-Syntetos-Babai (TSB) method. We include TSB as another benchmark. Single exponential smoothing (SES) has also been applied to intermittent time series with some success (Wallström and Segerstedt, 2010) and we include this benchmark as well. Note that although Croston's method and its variants are often expected to do better, there are combinations of demand and interval variability where SES performs satisfactory, in particular when the observed intermittency is low (Syntetos et al., 2005; Petropoulos and Kourentzes, 2014). Kourentzes (2014) discusses the impact of different cost functions on the parameter optimization of Croston's method and its variants and proposes two alternatives based on the notion of demand rate that perform better than conventional mean squared (or absolute) error. We exploit his results in the definition of our proposed cost functions.

In the literature other approaches to deal with this problem have been investigated: Zhu (2006) looks specifically at unconstraining for car rentals by exploiting turndown information. The approach requires the ability to identify customers so as to determine duplicates or re-books, which one often does not have. Other RM literature on car rentals discusses optimization approaches. Schmidt (2009) formulates the car rental booking control problem into a Markov decision process and proposes a few linear programming approximations to develop booking limit policies. Haensel et al. (2012) address a similar problem with stochastic programming (SP). Moreover, Steinhardt and Gönsch (2012) use dynamic programming to plan upgrades, while Su et al. (2012) consider downgrading decisions in a scenario of a firm providing two-class services where the low-class service can be used to accommodate high-class customers in exchange for a discount. Li and Pang (2017) propose a decomposition approach to address fleet movement between rental stations. In our simulation of a RM system, we use a probabilistic non-linear program motivated by the actual optimization module implemented at our industrial partner.

3 Unconstraining Methods

In this section we discuss our proposed methods for unconstraining demand along with benchmarks. We separate them into two major classes, those that are modelled on data presented in the form of: (i) *Booking Curves*; and (ii) *Booking Arrivals*. Both represent the same data. Booking curves show the cumulative sold units whilst booking arrivals show the arrivals of individual sales events (see Figure 1). The different representations are important in as far as different forecasting methods are used to unconstrain demand for each.

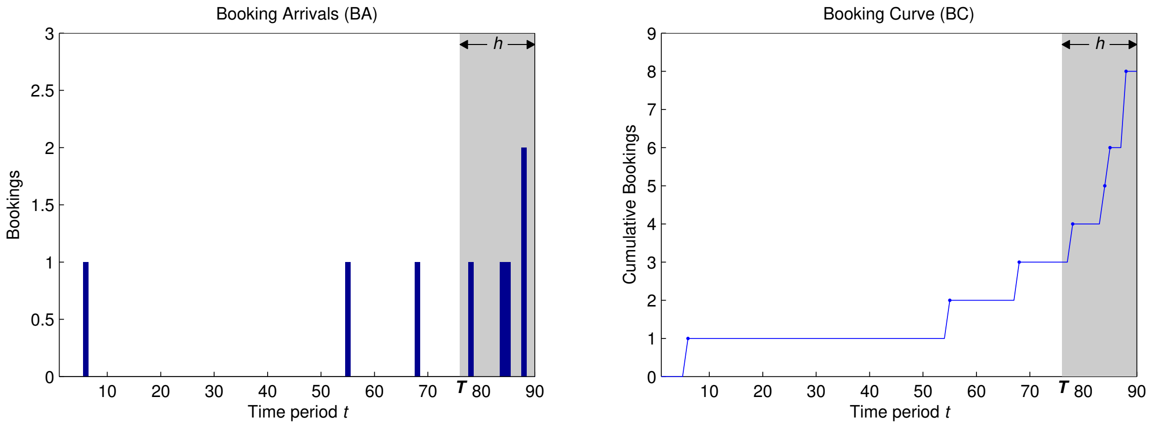

Figure 1: Illustration of Booking Arrivals (BA) and Booking Curve (BC) for a car rental product. In period T a booking restriction is introduced h periods prior to pick-up.

Observe in Figure 1 that a part of the demand is occurring after the booking restriction is imposed. If this is not considered then the true demand will not be accounted during decision making, leading to suboptimal results. Unconstraining methods are tasked to meet this unseen demand. We are particularly interested in unconstraining data with infrequent booking events. The rate of arrivals over time determines the shape of the booking curve; typically it is either (i) homogeneous, when the rate of increase is constant; (ii) concave, when the rate of increase is decreasing; or (iii) convex, when the rate of increase is increasing (Liu et al., 2002). Convex curves are typical for short-haul flights, business hotels or urban car rental stations. Concave ones are often found for e.g. at resort hotels where customers book long in advance. Of course, the notion of convexity/concavity in this context is not to be understood in the strict mathematical sense. The classic intermittent demand case as found in the forecasting literature corresponds only to homogeneous booking curves. This raises a question as to the applicability of conventional intermittent demand methods, which we discuss in the following sections.

Hereafter, unconstraining methods are classified into two categories, those fitted on booking curves (BC methods for short) and those fitted on booking arrivals (BA methods). For the sake of simplicity, we assume that for any product at most one restriction is imposed at some time period T (on a uniform time grid) and that it remains in force until the end of the booking horizon. We denote the number of remaining time periods at the time of restriction until the end of the booking horizon by $h \geq 0$. We index time forwards, i.e. the booking horizon begins in time period $t = 1$. If a booking horizon is not constrained, then T denotes the end of the horizon and $h = 0$. The generalization to multiple time intervals of imposing and lifting restrictions, in our context, is straightforward. First, the unconstraining method is used to predict the demand for the duration of a restriction, subsequently this prediction is added to the restricted booking curve for the duration of the restriction and the process is repeated, until all restriction periods are unconstrained.

3.1 Unconstraining Booking Curves

Holt's method was investigated and found to perform very well by Queenan et al. (2007). Building on this, we propose to use the *Damped Trend* model, a refinement of Holt's method. We compare the two approaches with benchmark methods specified in the online supplement, namely the *Averaging* method proposed by Weatherford and Pölt (2002) and the random walk, also sometimes called the naive forecast.

Let us briefly recap Holt's method as proposed by Queenan et al. (2007). The advantage of this linear trend exponential smoothing approach is that it takes into account when a product was constrained over the booking horizon. The idea behind this is to consider the booking curve as a trending time series. Any observations before the restriction are used to fit the Holt's method that models the series as a pair of dynamic level and trend components. Based on these, a forecast is produced, extrapolating linearly the estimated trend in the booking curve to produce a projected unconstrained cumulative demand figure. Although Holt's method can only produce linear trend forecasts, it can capture stochastic trends (i.e. trends whose rate of change can evolve over time, see b_t in (1)) thus being a good candidate to model a wide variety of booking curves. Hyndman et al. (2002) embedded Holt's method within the ETS (ExponenTial Smoothing) state space model, providing the statistical rationale behind it and deriving the likelihood function that permits identifying optimal parameters. The formulation of the model is:

$$\begin{aligned} F_{T+h} &= l_T + hb_T, \\ l_t &= l_{t-1} + b_{t-1} + \alpha e_{t-1}, \\ b_t &= b_{t-1} + \beta e_{t-1}. \end{aligned} \tag{1}$$

The model separates the observed actuals A_t, the value of the booking curve at period t, into a level l_t and trend component b_t, which are updated in each period by $e_t = A_t - F_t$, factored by the smoothing parameters α and β. These smoothing parameters can take values between 0 and 1 and can be interpreted as the percentage that each component is updated based on the last prediction error. Higher parameters result in more reactive model fit, which on the other hand may not filter adequately the noise in the observed data. Both smoothing parameters and the initial values l_0 and b_0 can be estimated by maximum likelihood. Because this model assumes additive errors this is equivalent to minimizing the in-sample Mean Squared Error (MSE):

$$MSE = T^{-1} \sum_{t=1}^{T} e_t^2, \tag{2}$$

where T is the number of observations in-sample, i.e. the periods up to the booking restriction.

Although linear trend ETS has been shown to perform very well empirically for unconstraining demand (Queenan et al., 2007), it is appropriate only for homogeneous booking curves, as concave or convex curves are not satisfied by its linear trend extrapolation. This limitation can be overcome by extending the model to nonlinear trends, as we propose below. Naturally, over short time intervals a linear trend could be a reasonable approximation of nonlinear ones. However, the model misspecification will influence the quality of parameter estimation and in turn the quality of the forecasted trend.

Within the ETS family of models the *Damped Trend* ETS permits to model and forecast time series that match concave booking curves. This model is an extension of the linear trend one, where a new parameter ϕ is introduced to control for the amount of dampening that is applied on the linear trend, thus transforming it into a nonlinear trend:

$$\begin{aligned} F_{T+h} &= l_T + b_T \sum_{i=1}^{h} \phi^i, \\ l_t &= l_{t-1} + \phi b_{t-1} + \alpha e_{t-1}, \\ b_t &= \phi b_{t-1} + \beta e_{t-1}. \end{aligned} \tag{3}$$

If $\phi < 1$ the linear trend is damped, making it appropriate for concave booking curves. Note that if $\phi = 1$ then the damped trend model becomes equivalent to the linear trend ETS, thus being able to model homogeneous booking curves. Allowing $\phi > 1$ is appropriate to model convex booking curves;

this is in contrast to the forecasting literature that considers only $\phi \leq 1$ as meaningful: see e.g. Gardner and McKenzie (2011). Therefore, the Damped Trend model is theoretically appropriate to model all concave, homogeneous and convex booking curves. We anticipate that this refinement of Holt's method should exhibit better performance due to its increased flexibility. Its parameters and initial values can be estimated in the same way as was described for the Holt's method.

3.2 Unconstraining Booking Arrivals

We investigate the use of *Croston's method* with a new way of optimizing its parameters. As benchmarks, we use the Teunter-Syntetos-Boylan method (*TSB*) and *Single Exponential Smoothing.* Both benchmarks are defined in detail in the online supplement. All these methods have been identified as appropriate to model intermittent data in the forecasting literature, which corresponds to small demand in our context; hence we investigate their usability for the task of unconstraining demand.

Queenan et al. (2007) claimed that if the observed demand is small and intermittent, Croston's method (Croston, 1972) is more appropriate to use than other methods, as alternative methods will not be able to capture the observed booking curve dynamics accurately due to the intermittency of the data. This echoes the discussion in the forecasting literature between fast and slow moving items, where the intermittent demand of the latter requires special forecasting models, with Croston's method being the most widely used (Syntetos and Boylan, 2005).

The principal idea behind Croston's method is the following: since intermittent time series have variability both in the demand size and timing, we address each one separately. This is done by separating any intermittent demand time series into two components, a vector of non-zero demand observations z, and a vector of inter-demand intervals x. Each is then modelled independently, and the final prediction of Croston's method is the arrival rate expressed by the ratio of their predicted values:

$$f_{T+h} = \frac{\hat{z}_J}{\hat{x}_J}, \tag{4}$$

where f_{T+h} is the forecasted value of the booking arrivals series and J denotes the last observed arrival event. To use that to unconstrain a booking curve we need to consider the cumulative prediction:

$$F_{T+h} = A_T + \sum_{i=1}^{h} \frac{\hat{z}_J}{\hat{x}_J} = A_T + h\frac{\hat{z}_J}{\hat{x}_J}. \tag{5}$$

Therefore, Croston's method assumes a linear trend and thus can be expected to generally not work well on non-homogeneous booking curves.

Each of the two components in Croston's method z_J and x_J are modelled using simple exponential smoothing:

$$\hat{z}_j = \alpha_z D_j + (1 - \alpha_z) z_{j-1}, \tag{6}$$
$$\hat{x}_j = \alpha_x I_j + (1 - \alpha_x) x_{j-1},$$

where D_j and I_j are the observed non-zero demand and inter-demand interval at the j^{th} arrival event, while α_z and α_x are their respective smoothing parameters. Snyder (2002) discuss the advantages of not assuming equal α_z and α_x and Kourentzes (2014) provides evidence that this results in better performance, in contrast to the standard implementation of using a single parameter (as for instance in Queenan et al., 2007). Note that we use j index instead of time t as any periods with zero demand are not considered, a limitation that TSB overcomes.

We now introduce a new cost function for the parameter estimation of the methods defined on booking arrival data. Queenan et al. (2007) suggested to use the mean squared error (MSE) as a cost function, however Kourentzes (2014) demonstrated in an inventory setting that conventional errors, such as MSE, are not ideal for estimating intermittent demand method parameters.

Motivated by that work, we propose to use the Curve Mean Squared Error (CMSE) as cost function:

$$CMSE = T^{-1} \sum_{t=1}^{T} \left(A_t - \sum_{k=1}^{t} f_k \right)^2, \tag{7}$$

where A_t represents the actual cumulative demand up to and including the t^{th} time period, and f_k is the booking rate at time period k defined by the most recent rate update at arrival event j, that means $f_k := f_j = \hat{z}_j / \hat{x}_j$. Therefore, these errors are calculated between the cumulative booking arrivals, i.e. the booking curve, and the cumulative forecasts. This way we avoid comparing a 'booking rate' forecast as resulting from equation (4) to the observed booking arrivals; these are not comparable. The difference between the proposed CMSE and the cost functions by Kourentzes (2014) is that the former optimizes the data on a higher level of cumulation.

3.3 Group Parameter Estimation

For cases where very limited booking events are recorded, the number of these events in comparison to the number of parameters to be estimated becomes important. With very few points the estimated parameters will potentially produce very inaccurate demand unconstraining.

In this situation we propose to estimate the parameters across multiple booking curves of the same product simultaneously, so as to increase the available sample. By "the same product", we mean products that are the same in every aspect except the service delivery date; for example: all historic booking curves for rental of a particular car type, station, length-of-rental, for pick up on a specific weekday. We assume that the corresponding booking curves follow similar dynamics. We construct cost functions on the pooled errors across multiple series, thus having a large number of booking events. The MSE and CMSE cost functions described before can be reformulated as Group MSE and Group CMSE as follows:

$$GMSE = \sum_{m=1}^{M} MSE_m = \sum_{m=1}^{M} \left(T_m^{-1} \sum_{t=1}^{T_m} e_{m,t}^2 \right), \tag{8}$$

$$GCMSE = \sum_{m=1}^{M} CMSE_m = \sum_{m=1}^{M} \left(T_m^{-1} \sum_{t=1}^{T_m} \left(A_{m,t} - \sum_{k=1}^{t} f_{m,k} \right)^2 \right), \tag{9}$$

where M is the total number of booking curves grouped together, T_m the time period when the restriction is imposed on the m^{th} curve (or otherwise the end of the time horizon), $e_{m,t}$, $A_{m,t}$ and $f_{m,t}$ the errors, the cumulative booking arrivals actuals and forecasts for each booking curve and period respectively. To calculate $e_{m,t}$ the actual and predicted values for the m^{th} booking curve at time t are used, in analogy to (2).

4 Numerical Experiments

We evaluate the performance of the various methods in terms of accurately predicting unconstrained demand and in terms of the resulting improvement in revenue. To that end, we teamed up with a major car rental company in the United Kingdom to evaluate our methods and generated demand trajectories using distribution parameters estimated from actual car rental data. We build a somewhat simplified version of a full RM system that mirrors the functions of the system in place at our partner company. Using this system we are able to demonstrate revenue impact of the improved demand estimates.

As mentioned earlier, car rental is an interesting application for small demand unconstraining methods because they typically observe few bookings per product even at larger stations. Booking restrictions are implemented on the product level so that we cannot easily merge booking histories without loosing the information on when restrictions came into effect.

4.1 Description of Industry Data

The data was collected for their station at Heathrow Airport in London and corresponds to their busiest station in the country. It encompasses all bookings for pick-up dates in 2011, in total around 130,000 booking requests and features information on booking timestamps, pick-up and drop-off stations and timestamps, and car group. Another data set contains information on restrictions at the Heathrow station for the same time horizon, including timestamps of restriction start and end for what product (i.e. car group, pick-up date, length-of-rental).

We now outline the characteristics of the data that motivated the simulation design for the numerical experiments.

- *Advance booking pattern*: Customers may make their bookings as early as one year in advance, but most bookings arrive shortly prior to pick-up, namely 16% within 1 day to pick-up, 47% within 7 days, 83% within 30 days, and 95% within 90 days, hence the booking curves are typically convex. We use a 90 day horizon in the accuracy study.

- *Independent demand*: We assume products (i.e. the combination of station, car group, pick up date, length-of-rental) to be independent of each other. The company believes this to be a reasonable assumption since most customers are not loyal and they easily switch to another supplier if their desired car group is not available, especially for the Heathrow airport station where all major competitors are represented. Therefore, we do not consider demand spill and recapture. We limit the simulations to a single car group.

- *Demand size*: Total demand per product varies depending on car group, length-of-rental, pick-up date, etc. Typically, it falls in the range between 6 and 18 for the most popular car groups.

- *Restrictions*: The company uses product-level booking limit control and recorded when restrictions were imposed. Most of them came into effect shortly prior to pick-up, e.g. 12% one day in advance, 57% during the last 7 days prior to pick-up and about 90% within 30 days. Once in place, restrictions usually remained until the end of the booking horizon. Accordingly, we generate only at most one restriction per booking history lasting throughout the remaining time horizon. Unconstrained actual demand is not known when a restriction was in place.

- *Cancellations*: We assume in our simulation that no bookings are cancelled so as to simplify the analysis.

Figure 2 shows a few historical booking curves at London Heathrow Airport. It is obvious that they were all convex and their demand size was small. Most of the bookings arrived within 30 days prior to pick-up. In particular, booking curve 1 was unconstrained and thus bookings were observed until the pick-up date, while booking curve 2 and 3 were constrained 7 and 3 days prior to pick-up, respectively. We describe the generation of simulated booking curves and restriction times based on the actual car rental data in the online supplement.

In the following section, we present the numerical results on accuracy of various unconstraining methods. We summarize those methods in Table 1 for reference, in which our proposed methods are highlighted.

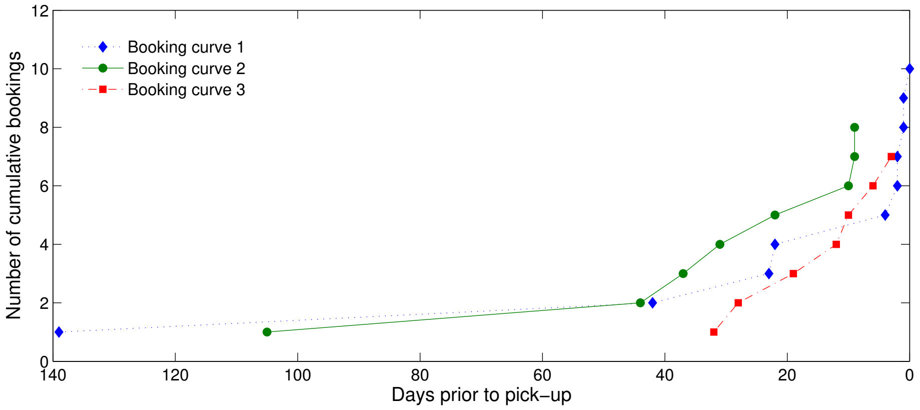

Figure 2: Sample booking curves at London Heathrow Airport

Table 1: A summary of unconstraining methods

Category	Method	Definition
BA methods	*Croston-MSE*	Croston's method as in (6) with MSE cost function (2)
	Croston-CMSE	Proposed variant of Croston's method as in (6) with CMSE cost function (7)
	TSB	Teunter-Syntetos-Babai approach as defined by (A.2) in the online supplement
	SES	Single ETS model as defined in the online supplement
BC methods	*Holt*	Holt's ETS model as defined by (1) with MSE cost function (2)
	Damped	Proposed Damped trend ETS model (3) with cost function MSE (2)
	Naive	As defined by (A.1) in the online supplement
	Averaging	As defined in the online supplement

4.2 Results on Accuracy

As we generate true booking curves over the entire time horizon and then add restrictions so as to truncate the data retrospectively, this enables us to measure the accuracy of each unconstraining method. We generate an estimate of unconstrained demand and compare it to the true total demand by measuring the Absolute Percentage Error between the real unconstrained demand and the predicted unconstrained demand, at the pick-up period for each booking curve:

$$APE_m = 100\frac{|A_m - F_m|}{A_m}. \tag{10}$$

These are then summarized across all booking curves to form the Mean Absolute Percentage Error (MAPE) and the Median Absolute Percentage Error (MdAPE). The reported figures in this section are the average MAPE and MdAPE across all 100 simulations (with 100 booking curves each). Due to space restrictions we report only the results for convex booking curves which are most prevalent in our car rental booking data. The results for concave curves are similar, while for homogeneous ones the performance of Croston-CMSE and Holt improves and becomes comparable to Damped, as the linear trend assumption holds in this case.

We measure the accuracy of unconstraining for booking curves with more than 4 booking events separately from those with less than that. We refer to the first group as booking curves with normal sample size, in the context of our application, and the latter group as booking curves with limited sample size, for which parametrization becomes very challenging. For the second case we provide only results for the grouped optimization. We first compare the impact of the new cost function on unconstraining accuracy for methods modelled on booking arrival data, then compare the accuracy of the various approaches (modelled either on booking arrivals or booking curves), test for statistical significance of the observed differences and finally investigate the impact of different levels of available unconstrained observations on the methods' accuracy.

Comparison of cost functions: We provide results for the performance of the proposed CMSE cost function against the benchmark MSE for the methods modelled on booking arrival data. Table 2 presents the MAPE and MdAPE for the three relevant methods, as well as the percentage improvements of CMSE over MSE.

Table 2: Unconstraining accuracy for BA methods, with MSE and CMSE cost functions

Cost Function	MAPE (in %)			MdAPE (in %)		
	Croston	TSB	SES	Croston	TSB	SES
MSE	38.2	38.3	43.2	37.4	38.3	43.5
CMSE	28.4	32.8	33.3	25.0	27.8	28.2
Improvement	+25.7%	+14.4%	+23.0%	+33.1%	+27.4%	+35.2%

For all BA methods the proposed CMSE cost function consistently performs better. Hereafter all results reported for the BA methods will be based on CMSE and its GCMSE counterpart.

Assessment of accuracy: We compare the accuracy of all methods and present the results in Table 3. These are organized as follows: first, they are separated depending on the number of booking events before the booking restriction is enforced; second they are separated between methods optimized on individual booking curves (Ind.) or at a group level (Group); third MAPE and MdAPE are reported. Each column represents one combination of the above categories and the best performing method is highlighted in boldface. The results are visualized to facilitate comparisons in Figure 3. The horizontal dashed lines represent the performance of the Naive, which represents a minimum performance threshold for any method to be considered acceptable, given additional complexity over the Naive.

Focusing on the booking curves with 'normal' sample size, overall MAPE and MdAPE exhibit similar behavior. The conclusion by Queenan et al. (2007) that Holt, optimized by MSE, is a good performer is validated in our results, outperforming both Naive and Averaging benchmarks. Croston-CMSE performs closely to Holt and better than other alternative methods fitted on the booking arrivals data.

Let us focus on the results obtained when the optimization of the methods is done on individual time series. Holt is marginally better than Croston-CMSE, while the opposite is true when looking at MdAPE. This can be explained by the presence of outliers in the distribution of the errors. Nonetheless, the differences in both cases are marginal. Under both measures Damped performs best, which is in-line

Table 3: Unconstraining accuracy

Method	Normal sample size				Limited sample size	
	MAPE (in %)		MdAPE (in %)		MAPE (in %)	MdAPE (in %)
	Ind.	Group	Ind.	Group	Group	Group
Croston-CMSE	28.4	32.0	25.0	30.6	63.4	64.1
TSB	32.8	33.6	27.8	28.4	89.9	85.2
SES	33.3	34.0	28.2	28.7	90.3	85.6
Holt	28.0	26.9	25.0	24.2	73.3	75.2
Damped	**26.4**	**21.0**	**21.4**	**18.0**	**44.8**	**44.3**
Naive	37.0	37.0	35.8	35.8	84.2	85.2
Averaging	28.8	28.8	27.2	27.2	70.5	72.2

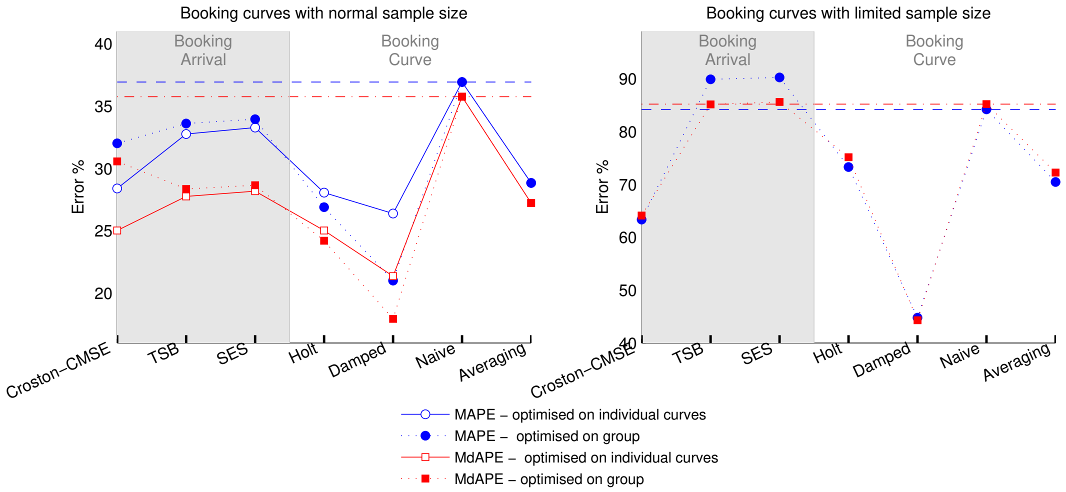

Figure 3: Unconstraining accuracy for booking curves with normal and limited sample size.

with the expected improvements over Holt given its flexibility to model convex booking curves. Therefore, Damped not only models a wider range of cases theoretically, but we also find empirical evidence that it is more accurate in unconstraining the demand for the non-homogeneous case where the linearity assumption is violated.

When looking at the performance of the methods optimized across a group of multiple booking curves, BC methods improve further, while BA methods perform on average marginally worse, effectively being always outperformed by Averaging. It is interesting to observe that the performance of Holt improves further, making it now substantially more accurate than Croston-CMSE in this case (note that Croston-CMSE is optimized in this case using GCMSE). Damped still performs best and the GMSE results are better than those based on individual optimization (MSE), exhibiting the overall best performance across all setups. Therefore, the results support the use of the proposed cost function to optimize the methods.

Turning our attention to time series with very limited number of booking events, results are provided only for the group optimization cost functions. Croston-CMSE is a good performing method and a good alternative to Holt, as reported by Queenan et al. (2007). Damped still performs best, in fact substantially better than the other benchmarks.

Statistical significance: Furthermore we explore whether the observed differences are significant or not, especially in the cases where the accuracy differences are small. To do this we use the nonparametric Friedman and post-hoc Nemenyi tests that do not impose any distributional assumptions (Demšar, 2006). Three comparisons are conducted: (i) booking curves with 'normal' sample size and methods optimized on each curve individually; or (ii) using group cost functions; and (iii) for booking curves with 'limited' sample size and optimized using group cost functions. In all scenarios Damped is found to perform significantly better than the rest of the methods.

Degree of constraints on the data: Finally, we investigate how the methods compare when applied to data that is highly constrained as opposed to data that is only slightly constrained. Figure 4 presents the results. When methods are optimized on individual booking curves the errors increase for higher percentages of constrained periods, as expected. It is also apparent that Croston-CMSE, Holt and Averaging do not differ substantially echoing the results of the Nemenyi test. However, when methods are optimized using the group cost functions a different behavior emerges, with Holt now performing substantially better than Averaging. Note that there is no difference in the results for Averaging under individual or group optimization, as there are no parameters to estimate. The best performing method remains to be Damped, which under this scenario is substantially better than all other alternatives and better in comparison to individually optimized Damped results. Looking at the results for booking curves with 'limited' sample the observed errors are higher, as expected. Damped still performs best and although its performance deteriorates rapidly as the percentage of constrained periods increases, in all cases it is 20-60% better than Croston-CMSE, the second most accurate method.

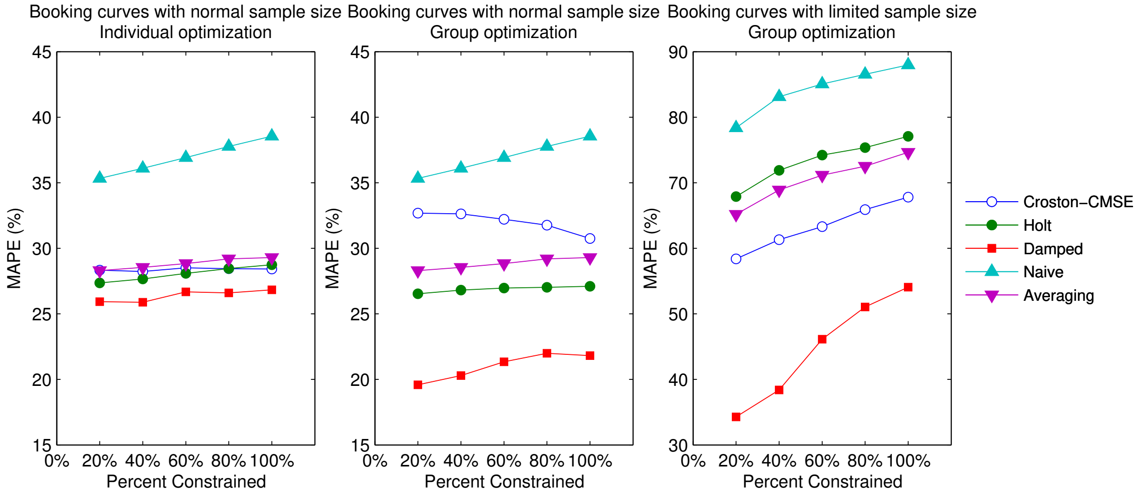

Figure 4: Accuracy for best performing methods (optimized on groups) for different restriction scenarios.

In conclusion, our empirical evaluation suggests that Damped, optimized on GMSE, performs best in all cases. The improvements observed are significant over benchmark methods from the literature and refinements proposed here, such as Holt that is found to improve when the proposed GMSE is used to identify optimal parameters.

4.3 Results on Revenue Impact

The ultimate aim of unconstraining is of course to improve the revenue performance of a RM system. Better unconstraining methods can lead to revenue gains, e.g. 0.5-1% improvements were quoted by Weatherford and Pölt (2002). We are interested in the extent to which our proposed methods lead to improved revenue over the benchmark methods. To that end, we replicate a somewhat simplified version of the RM system found at our car rental partner.

The framework of the simulation is depicted in Figure 5. The true demand booking trajectories for all products are generated *a priori*. They are then fed to a booking system which determines whether to accept or reject a particular customer booking, based on booking limits produced by the optimization module and on inventory availability. If accepted, the booking system registers the revenue and reduces the inventory level by one for the corresponding rental days. Rejected sales are lost. The output of the booking system is a history of observed sales records, as well as a history of when restrictions were in place, if any. The sales records are used to forecast future demand if no unconstraining is carried out.

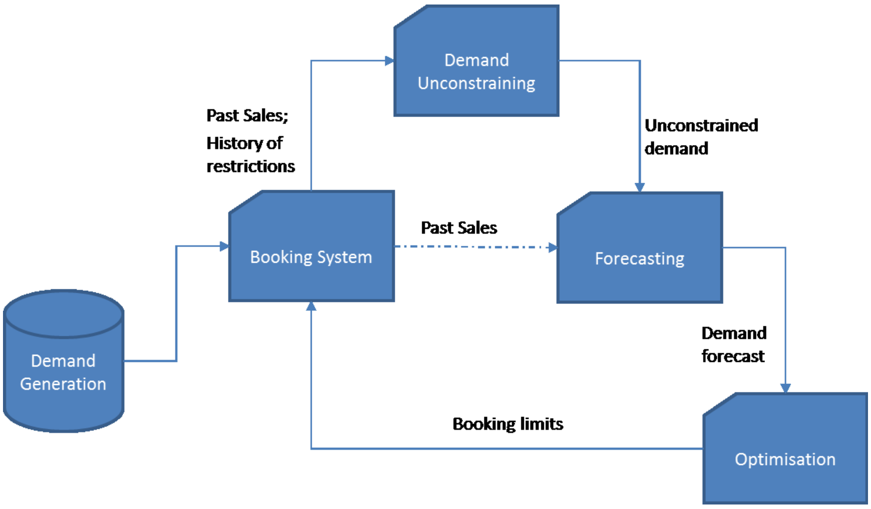

Figure 5: Rental car revenue management simulation framework. See the online supplement for detailed description.

Otherwise, they are passed to the demand unconstraining module to estimate the true market demand to be used in the forecasting module. We include in the simulation's unconstraining module those strong performing methods in the accuracy evaluation: Croston's method with MSE-based cost function (Croston-MSE), Croston's method with CMSE-based cost function (Croston-CMSE), Holt's method (Holt), and damped Holt's method (Damped). Note that the parameters in all these methods are optimized at group level. We also include two benchmarks, i.e., first come first served (FCFS), and no unconstraining (i.e. using constrained demand in the forecast (Constr)).

The optimization module determines the booking limits given the forecast of demand-to-come and the remaining inventory levels. The booking limit control is then used by the booking system. The forecasting and optimization modules run once a day so as to update the booking limits according to observed sales.

We make the same assumptions on the demand generation process as discussed in Section 4.1, and shortened the advanced booking horizon to 30 days which seemed reasonable given that 83% of actual bookings in the data were received within this time horizon. Moreover, each booking day is split into three time periods, which allows a more granular description of the booking curves. Such a treatment also allows restrictions to be implemented within a day, which is not uncommon in practice. In the revenue simulations, we consider products for a fixed car group with length-of-rental of at most 7 days because we observed nearly 80% of bookings are for 7 day rentals or less. Customers for such short rentals usually book relatively shortly prior to pick-up, while those for long rentals will book long in advance to secure their cars. This emphasizes the need for unconstraining methods to be able to deal with both convex and concave booking curves. We account for seasonality effects since car rental demand has a distinct weekly pattern. For airport stations, more demand is seen in weekdays and less in weekends due to business travellers, especially on Mondays. A more detailed description of the simulated RM system and data generation can be found in the online supplement.

We conduct 100 simulation runs for each unconstraining method. In each simulation run, the total revenue obtained over the considered time horizon (excluding a warmup period) is compared to the perfect information benchmark in which the true demand is known in advance. All revenue results are presented as the percentage loss to this perfect information scenario. We consider between 180 and 360 pick-up dates depending on the length of the simulation's time horizon. The rental rates are typical retail prices for the considered car group in the case company. For simplicity we assume the rental rates only dependent upon the length-of-rental. Following the actual pricing practice, the longer the length-of-rental, the lower the daily rate.

In the following, we first discuss the results of revenue impact of unconstraining over time for a fixed fleet size and, secondly, consider revenue impact of unconstraining over fleet sizes for a fixed simulation time horizon.

Revenue impact of unconstraining over time: The percentage revenue losses are reported in Table 4 and plotted in Figure 6, which also includes the unconstraining error as measured using MAPE. The fleet size is fixed at 140. The spiral-down effect can be observed in the performance of Constr, namely that using constrained demand leads to declining revenues over time. It is interesting to observe that using no RM system at all (i.e. using FCFS) is much better than to use constrained demand in the considered RM system.

All unconstraining algorithms' revenue performance improves over time. Our proposed methods Croston-CMSE and Damped are consistently producing the best revenue results with improvements over

Table 4: Mean percentage revenue loss over different time horizon, with fixed fleet 140.

Horizon	180	210	240	270	300	330	360
Constr	12.8	13.0	13.7	14.4	15.2	15.8	16.4
FCFS	12.1	11.1	11.0	10.9	10.8	10.7	10.7
Holt	8.7	7.8	7.4	7.1	6.9	6.8	6.7
Croston-MSE	8.4	7.6	7.1	6.8	6.4	6.3	6.0
Damped	7.3	6.7	6.4	6.2	6.0	5.9	5.8
Croston-CMSE	7.3	6.7	6.3	6.0	5.8	5.7	5.5

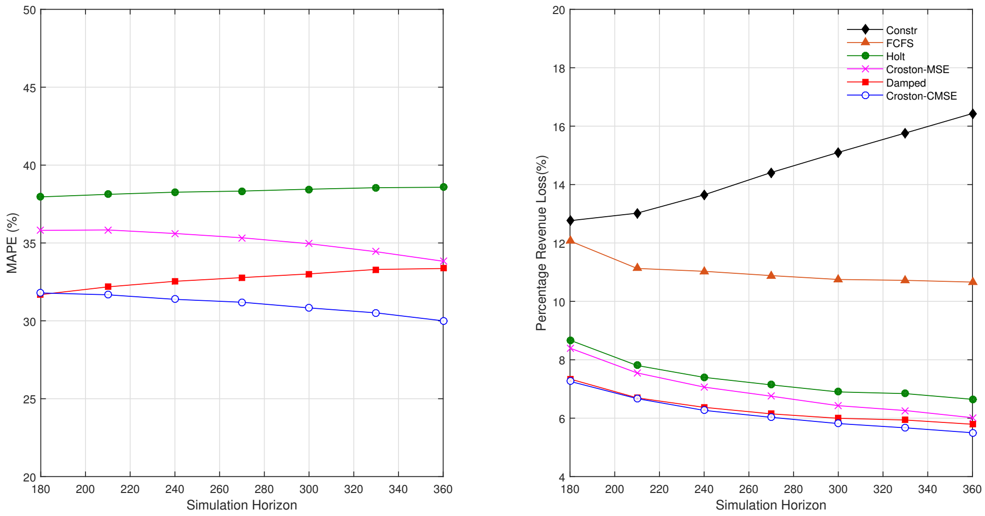

Figure 6: MAPE and mean percentage revenue loss over time, with fixed fleet 140.

Constr in the range of 5.5% to 10.9%. They also improve by 0.5%–1.1% over Croston-MSE and by 1.1%–1.4% over Holt, respectively. The two proposed methods produce overall similar revenue results.

To understand whether the revenue improvements are statistically significant, we calculated the 95% confidence intervals around the observed mean revenue for each of the four unconstraining algorithms. In all scenarios there is no overlapping of the confidence intervals between our proposed methods and the benchmarks, and thus the revenue improvement is significant.

In terms of the unconstraining accuracy, our results in Figure 6 show that Croston-CMSE is the best alternative for all simulation time horizons. This is in contrast to our accuracy analysis that found Damped to be the best, which was only outperformed by Croston-CMSE for small sample sizes and highly constrained scenarios. To further understand this we calculated MAPE by LoR for all methods. The results for selected simulation horizons are plotted in Figure 7. We find that all algorithms' performances deteriorate with LoR, which is not surprising as the demand size for longer rentals is much smaller than their shorter counterparts. For every simulation horizon the best unconstraining algorithm is Damped for shorter LoRs while for longer rentals Croston-CMSE becomes the one. We have found that the restriction level is particularly high especially for longer LoR cases, where usually there is not enough data to identify clear nonlinear booking patterns. Therefore Damped's performance is expected to deteriorate. Even though the performance of the other three algorithms also reduces with LoR, their reduction rates are slower.

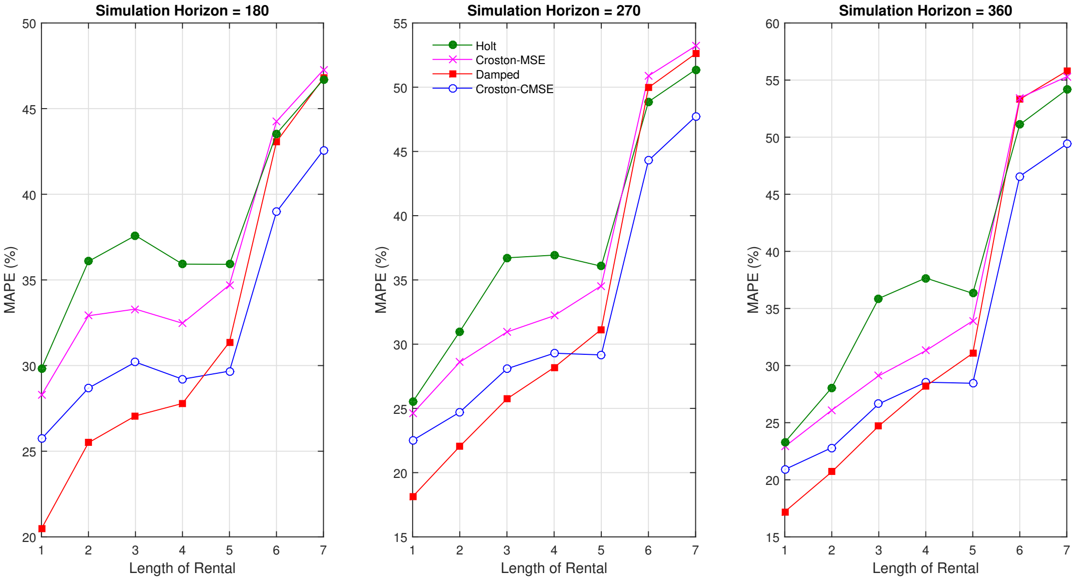

Figure 7: MAPE versus length of rentals for selected simulation horizons, with fleet size 140.

For shorter rentals the accuracy of Holt and Damped improved with time. However, their performance degrades for longer rentals due to limited sample sizes. In contrast, even though both Croston's methods also experience reduced performance over time for longer LoRs, the drop is much more moderate. This explains why Figure 6 shows that the accuracy of Holt and Damped reduces while that of both Croston's methods improves with time.

In Figure 6, it is shown that even though the unconstraining accuracy reduces with the simulation horizon for Damped and Holt, their revenue performance still improves. This perhaps somewhat surprising result is actually due to the unbalanced demand volume across LoRs, with higher demand for shorter ones. Therefore, the revenue contribution is dominated by shorter rentals whose accuracy improves over time for both algorithms.

Revenue impact of unconstraining depending on fleet size: Capacity tightness, i.e. the ratio of average daily true demand over daily capacity, is an important factor to consider when evaluating revenue performance in simulations. Intuitively, if capacity by far exceeds demand, then the optimal policy is to accept all demand (FCFS). The other way around, if demand by far exceeds capacity, one should only accept one day rentals because the per-day rate is highest for them. A RM system will be most effective in the scenarios where the capacity is not enough to accommodate all demand, and it is essential to trade-off between multiple demands which usually arrive in different times.

We consider fleet sizes from 110 up to 200, increasing by 10. Figure 8 shows MAPE and the percentage revenue loss relative to perfect information scenario for different fleet sizes over a time horizon of 300 days. As expected, FCFS produces similar results like all unconstraining methods for large fleet sizes. It is not surprising that the performances of all unconstraining methods improve with the fleet size. Moreover, the differences between algorithms reduce as well with the fleet size. For the more interesting cases of tighter capacity as often found in practice, the average revenue differences between the best-performing method Croston-CMSE and the least-performing Holt are as large as 2.2%.

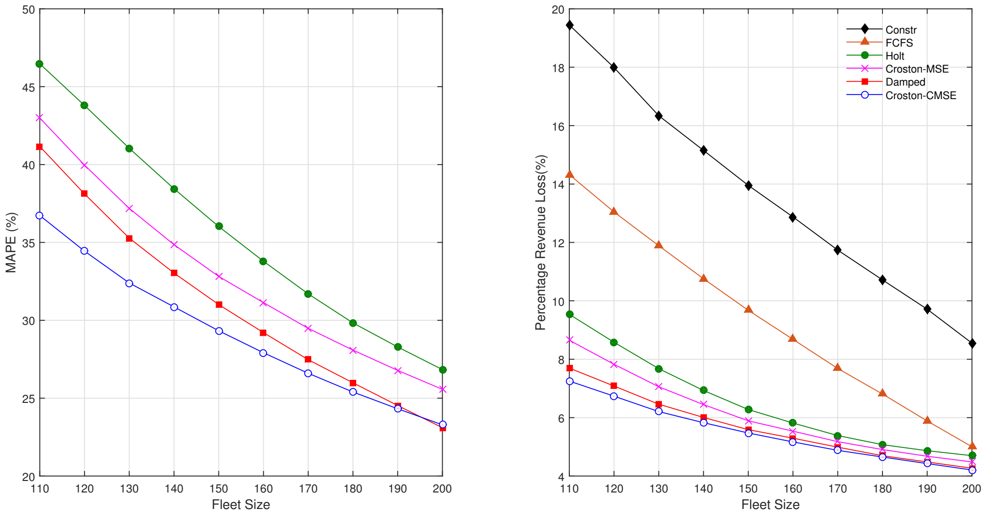

Figure 8: MAPE and mean percentage revenue loss for different fleet sizes over 300 days time horizon.

The unconstraining accuracy (MAPE) results are similar to those in the previous study: Croston-CMSE is the best alternative in most scenarios. However, its advantage over Damped diminishes over increased fleet size. This is easy to explain as larger fleet sizes allow Damped to capture the nonlinear trend better due to the increased number of booking curves that are used with the group cost function, even though each individual curve has limited sample size. For larger fleet sizes, the capacity is less tight and thus the booking curves less constrained. This also helps Damped to perform stronger. In all scenarios, our proposed methods statistically significantly improve the benchmarks at the 95% level.

5 Conclusion and Future Work

We propose to apply results from small and intermittent demand forecasting to unconstraining problems commonly encountered in revenue management. Specifically, we find that damped trend exponential smoothing and Croston's method with a special cost function can not only substantially reduce the estimation error with respect to true demand as compared to existing techniques, but also improve expected revenues by more than 1% over Holt's method. Revenue improvements on this magnitude are significant in revenue management applications because they translate directly into profits given that marginal costs are often close to zero. The simplicity and robustness of the proposed techniques is very appealing for industrial implementation.

The key take-away for modellers is that Croston-CMSE produces stronger results than Damped for sales data that is characterized by limited sample sizes and being highly constrained, and vice versa. Both methods produce better results than the benchmarks that we considered. It is common in RM practice that some products have lower demand volumes and/or are more constrained than the others, and thus a single method might not be able to accommodate all scenarios. If a product's sales data typically exhibits the same characteristics, the modeller could routinely use Croston-CMSE or Damped to unconstrain. If demand characteristics vary, we expect that an automated approach that switches between the two methods depending on some observable signal (such as the length of rental) will produce good results; we leave this to future research.

All of our discussed methods assume independence of demand, i.e. unavailability of a product is assumed to have no effect on the demand for other products. This assumption is the main limitations of our work since in various network RM applications this does not hold and may lead to double-counting of demand. However, it can be justified for applications where product alternatives are well fenced off, or where alternatives from competitors are readily available. Specific examples of the latter include car rental or casinos. If demand is dependent on availability of other alternatives, then substitution effects need to be taken into account.

Acknowledgement

All authors contributed equally and are named in alphabetical order.

Online Supplement

The online supplement can be accessed via the following link: http://eprints.whiterose.ac.uk/118673/.

References

Azadeh, S., Marcotte, P., Savard, G., 2014. A taxonomy of demand uncensoring methods in revenue management. Journal of Revenue and Pricing Management 13, 440–456.

Babai, M. Z., Ali, M. M., Nikolopoulos, K., 2012. Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis. Omega 40 (6), 713–721.

Bacchetti, A., Saccani, N., 2012. Spare parts classification and demand forecasting for stock control: Investigating the gap between research and practice. Omega 40 (6), 722–737.

Cooper, W., Homem-de Mello, T., Kleywegt, A., 2006. Models of the spiral-down effect in revenue management. Operations Research 54, 968–987.

Croston, J. D., 1972. Forecasting and stock control for intermittent demands. Operational Research Quarterly 23, 289–303.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30.

Gardner, E., 2006. Exponential smoothing: The state of the art – part II. International Journal of Forecasting 22, 637–666.

Gardner, E., McKenzie, E., 2011. Why the damped trend works. Journal of the Operational Research Society 62 (6), 1177–1180.

Gardner, E. S., J., McKenzie, E., 1985. Forecasting trends in time series. Management Science 31 (10), 1237–1246.

Guo, P., Xiao, B., Li, J., 2012. Unconstraining methods in revenue management systems: Research overview and prospects. Advances in Operations Research.

Haensel, A., Mederer, M., Schmidt, H., 2012. Revenue management in the car rental industry: A stochastic programming approach. J Revenue Pricing Manag 11 (1), 99–108.
URL `http://dx.doi.org/10.1057/rpm.2010.52`

Hyndman, R. J., Koehler, A. B., Snyder, R. D., Grose, S., 2002. A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting 18 (3), 439–454.

Kourentzes, N., 2014. On intermittent demand model optimisation and selection. International Journal of Production Economics 156, 180–190.

Li, D., Pang, Z., 2017. Dynamic booking control for car rental revenue management: A decomposition approach. European Journal of Operational Research 256 (3), 850–867.

Liu, P. H., Smith, S., Orkin, E. B., Carey, G., 2002. Estimating unconstrained hotel demand based on censored booking data. Journal of Revenue and pricing Management 1 (2), 121–138.

Petropoulos, F., Kourentzes, N., 2014. Forecast combinations for intermittent demand. Journal of the Operational Research Society 66 (6), 914–924.

Queenan, C., Ferguson, M., Higbie, J., Kapoor, R., 2007. A comparison of unconstraining methods to improve revenue management systems. Production and Operations Management 16, 729–746.

Schmidt, H., 2009. Simultaneous control of demand and supply in revenue management with flexible capacity. PhD thesis, Clausthal University of Technology, Germany.

Snyder, R., 2002. Forecasting sales of slow and fast moving inventories. European Journal of Operational Research 140 (3), 684–699.

Steinhardt, C., Gönsch, J., 2012. Integrated revenue management approaches for capacity control with planned upgrades. European Journal of Operational Research 223, 380–391.

Su, P., Tian, Z., Wang, H., 2012. On service degrade at a discount: Capacity , demand pooling, and optimal discounting. Omega 40, 358–367.

Syntetos, A., Boylan, J., Croston, J., 2005. On the categorization of demand patterns. Journal of the Operational Research Society 56 (5), 495–503.

Syntetos, A., Boylan, J. E., 2001. On the bias of intermittent demand estimates. International Journal of Production Economics 71 (1), 457–466.

Syntetos, A., Boylan, J. E., 2005. The accuracy of intermittent demand estimates. International Journal of forecasting 21 (2), 303–314.

Talluri, K. T., van Ryzin, G. J., 2006. The theory and practice of revenue management. Vol. 68. springer.

Teunter, R., Syntetos, A., Babai, M. Z., 2011. Intermittent demand: Linking forecasting to inventory obsolescence. European Journal of Operational Research 214, 606–615.

Wallström, P., Segerstedt, A., 2010. Evaluation of forecasting error measurements and techniques for intermittent demand. International Journal of Production Economics 128 (2), 625–636.

Weatherford, L. R., 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.

Willemain, T. R., Smart, C. N., Shockor, J. H., DeSautels, P. A., 1994. Forecasting intermittent demand in manufacturing: a comparative evaluation of crostons method. International Journal of Forecasting 10, 529–538.

Zhu, J., 2006. Using turndowns to estimate the latent demand in a car rental unconstrained demand forecast. Journal of Revenue and Pricing Management 4, 344–353.

Online Supplement to "Unconstraining Methods for Revenue Management Systems under Small Demand"

Nikolaos Kourentzes, Dong Li, Arne K. Strauss

A Benchmark Unconstraining Methods

Averaging

Weatherford and Pölt (2002) propose a simple averaging method to unconstrain data that can be applied to small and intermittent demand as well. For a given historical booking curve, we divide the time horizon into 10 equal-sized periods and classify each as open if no restriction existed throughout the entire period, or otherwise as constrained. We calculate the average demand received over the open periods. For each closed period, we define estimated demand as the maximum of the observed demand in that period (which may happen if it was only partially constrained) and the average demand over the open periods. The resulting unconstrained total is the sum over all 10 estimations.

Random Walk

Both exponential smoothing methods above are based on the assumption that the observed booking curve, up to the period that the restriction is enforced, contains useful time dynamics that can be modelled. The random walk model, also known as the Naive, would operate on the assumption that all the information is contained in the very last period, i.e. when the restriction is imposed and hence the unconstrained demand is:

$$F_{T+h} = A_T. \tag{A.1}$$

The random walk model has the advantage that it has no parameters to estimate and hence can be used in any circumstances, irrespective of data availability or how many bookings have occurred prior to the restriction period. As such, it can be used as a powerful benchmark for any more complex unconstraining demand methods. We argue that any more complex methods should outperform the naive.

Teunter-Synthetos-Babai Method

Teunter et al. (2011) recognized that a limitation of Croston's method is that it reacts very slowly to information, only when a demand is observed, and therefore does not update its estimate when there are long periods of zero demand. Motivated by an inventory setting, Teunter et al. (2011) argued that for items with long periods of inactivity modelling obsolescence is important and proposed a new Croston-type method for intermittent data. The Teunter-Synthetos-Babai (TSB) method separates the time series into two components, the non-zero demand (z_t) and the probability of a demand event (p_t). The non-zero demand is modelled in the same way as for Croston's method. The probability of a demand event is a vector that is equal to 0 when no demand was observed and equal to one otherwise. This vector is then modelled with single exponential smoothing, resulting in a predicted probability of demand for the future periods. Note that the the demand size estimate updates only when a demand is observed, while the probability of demand updates every period. The final forecast is:

$$f_{T+h} = \hat{z}_T \hat{p}_T. \tag{A.2}$$

When there are long periods of zero demand $\hat{p}_t$ becomes lower, reflecting the higher probability of obsolescence.

Single Exponential Smoothing

Furthermore, SES is a simpler model compared to other demand prediction methods, having half as many parameters. Hence, it requires less data to optimize which is desirable when dealing with sparse booking arrivals. In this context SES is used to model the booking arrivals series and an expected rate of booking arrivals is produced. This is then cumulated in the same way as it was described for Croston's method to unconstrain the demand of the booking curve.

B Data Generation for Accuracy Study

Generation of Booking Curves

We consider 90 days per booking curve with daily Poisson arrival rates shown in Table B1. The rates are defined in a way such that overall total expected demand equals the mean demand figure in the first row. The percentage split of demand over the four time windows is the same in every demand scenario, namely 17%, 36%, 31% and 16%, starting with window 90-30 and ending with period 1-0, respectively. This percentage split has been derived from the car rental data set.

Table B1: Poisson arrival rates for a fixed car group/length-of-rental.

	Mean demand			
Days to pick-up	6	10	14	18
90-30	0.02	0.03	0.04	0.05
30-7	0.09	0.16	0.22	0.28
7-1	0.31	0.52	0.72	0.93
1-0	0.96	1.60	2.24	2.88

We generate a collection of 100 booking curves from this non-homogeneous Poisson distribution for a given mean demand scenario, representing the collection of all available booking histories for the same car group/station/length-of-rental of the same pick-up weekday, say, Monday, during the same season (assuming that there is seasonality over the year).

Generation of Restrictions

To investigate the impact of different degrees of available unconstrained observations, we use the following approach proposed by Queenan et al. (2007): first, we assume that the true final demand is Poisson distributed with means as shown above, namely $\mu := 6, 10, 14$ or 18. Next, for each mean demand scenario, we determine a cutoff value that represents the demand level above that the cumulative Poisson probability sums up to 20%, 40%, 60%, 80% or 100%, respectively. For example, the cutoff value for a 20% restriction level for Poisson-distributed final demand with mean $\mu = 6$ would be 8, i.e. the inverse cumulative Poisson distribution evaluated at 0.8. For each generated booking curve whose final demand exceeds or equals the cutoff value, we subsequently sample a restriction start time from the empirical distribution of restriction start times in the actual data.

We assume that the restriction remains in place until the end of the booking horizon. This assumption is not a particularly strong one since most restrictions in the actual data indeed satisfied that assumption; this is not surprising given that restrictions typically were imposed only shortly prior to the pick-up date.

C Specification of the Simulated RM System

The setup of our RM system aims to replicate the key modules of the system in place at our collaboration partner. We describe the modules that comprise the RM system as shown in Fig 5 in the following.

Demand Generation

We consider demand for a single car group only. Each product represents a combination of pick-up date and length-of-rental. The true booking curve for each product is generated by an non-homogeneous Poisson process as described above. Accordingly, the arrival rates over the entire booking horizon resemble the actual booking curves observed in the airport rental station. Each booking day is divided into three booking segments and thus the 30 days booking horizon is discretized into 90 time periods. The percentage of demand arrivals in each booking period, or the booking curve, is summarised in Table C1. Note that they are different across LoR. Compared to Table B1, we have considered more granular booking processes in the simulation.

The weekly seasonality pattern is reflected by higher demand being generated for weekdays and less for weekends. The mean demand per day is calculated from the actual customer bookings for a particular car group, which are summarized in Tables C2. The distribution of demand for different length-of-rental products reflects the empirical distribution of the same normalized to cover up to 7 days rentals and is

Table C1: Percentage of bookings in each booking period.

Time periods	LoR						
to pick-up	1	2	3	4	5	6	7
90-63	0.002	0.003	0.003	0.004	0.005	0.005	0.005
62-42	0.004	0.005	0.005	0.006	0.007	0.007	0.007
41-21	0.010	0.009	0.010	0.011	0.011	0.011	0.009
20-6	0.020	0.021	0.022	0.021	0.021	0.018	0.017
5-3	0.037	0.035	0.030	0.025	0.023	0.026	0.031
2-0	0.075	0.068	0.057	0.049	0.041	0.044	0.056

reported in Table C3.

Table C2: Mean demand by day-of-week.

DoW	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Daily Demand	71	96	79	85	86	87	64

Table C3: Demand distribution over length-of-rental.

LoR	1	2	3	4	5	6	7
Percentage	0.25	0.21	0.18	0.14	0.10	0.06	0.07

For each product, demand arrivals within each booking period are generated randomly by a Poisson distribution with the arrival rate obtained for the specific day of week, length-of-rental and booking period. Within one time period there could be multiple arrivals for different products since we sample from a Poisson distribution for each product separately. In the situation of multiple arrivals, their order is determined by random permutation. Since we consider length-of-rentals of up to 7, we generate in total $7H$ booking curves in each simulation run, where H is the simulation's time horizon.

Unconstraining

In unconstraining a particular product, all the historical booking curves of the same day-of-week and length-of-rental are taken into account. Therefore the unconstraining accuracy improves with the simulation progressing as more history becomes available. An initial horizon of 60 days, which forms part of the warm-up phase, is processed without unconstraining so as to have sufficiently many historic booking curves.

Forecasting

Without unconstraining, the observed sales record is used for forecasting future demand; otherwise, we use the unconstrained demand estimates. We have chosen the moving average as forecasting method in the simulation study so as to eliminate further need for parameter optimization. Only the demand histories for the same day-of-week and length-of-rental are used in each forecast. The forecasting module simply uses the average over the corresponding demand estimates provided by the unconstraining module. If no unconstraining is used, it averages the constrained sales records.

Optimization

The RM system's optimization module uses the probabilistic non-linear program (PNLP) suggested for the car rental application by Schmidt (2009). Let $r_{s,l}$ denote the rental rate (price) for the product corresponding to pick up at day s and for length-of-rental l. The random variable $Y_{s,l}$ represents future demand for product (s,l) and $B_{s,l}$ is its booking limit. The fleet available at s is denoted by scalar A_s. The optimization horizon is $N = 30$ days. The PNLP at day t can be stated as follows:

$$\text{PNLP: } \max \sum_{s=t}^{t+N} \sum_{l=1}^{7} r_{s,l} u_{s,l} \tag{C.1}$$

$$\text{s.t. } u_{s,l} = \mathbb{E}\left[\min\{B_{s,l}, Y_{s,l}\}\right], \qquad \forall\, s \in \{t, \ldots, t+N\}, \forall\, l \in \{1, \ldots, 7\}, \tag{C.2}$$

$$\sum_{\tau=t}^{s} \sum_{l=s-\tau+1}^{7} B_{\tau,l} \leq A_s, \qquad \forall\, s \in \{t, \ldots, t+N\}. \tag{C.3}$$

Since the demand variable $Y_{s,l}$ is discrete, we can calculate (C.2) by

$$u_{s,l} = \mathbb{E}[\min\{B_{s,l}, Y_{s,l}\}] = B_{s,l} - \sum_{y=0}^{B_{s,l}-1} F_{s,l}(y), \tag{C.4}$$

where $F_{s,l}$ is the cdf function for $Y_{s,l}$. It is obvious that $u_{s,l}$ is an increasing and concave function of $B_{s,l}$. In light of this property and the finite bound for $B_{s,l}$, equation (C.2) can be approximated by a set of piecewise linear functions. Specifically, it can be replaced by the following constraints.

$$u_{s,l} \leq \alpha^i_{s,l} B_{s,l} + \beta^i_{s,l}, \tag{C.5}$$

where $\alpha^i_{s,l}, \beta^i_{s,l}$ are the parameters for the $i^{th}(1 \leq i \leq I)$ linear function for product (s,l).

We next present how to determine these parameters. For a comprehensive account on this process refer to Talluri and van Ryzin (2006). For each product (s,l), sample $I+1$ booking limit values in between 0 and A_s, denoted by $B^i_{s,l}$. Substitute each of them into equation (C.4) and denote the result by $u^i_{s,l}$. Essentially we have just calculated the expected demand to be accepted for $I+1$ booking limit values. These $I+1$ pairs of $(B^i_{s,l}, u^i_{s,l})$ determine I linear functions whose parameters are given by,

$$\alpha^i_{s,l} = \frac{u^{i+1}_{s,l} - u^i_{s,l}}{B^{i+1}_{s,l} - B^i_{s,l}}, \tag{C.6}$$

$$\beta^i_{s,l} = \frac{u^i_{s,l} B^{i+1}_{s,l} - u^{i+1}_{s,l} B^i_{s,l}}{B^{i+1}_{s,l} - B^i_{z,l}}. \tag{C.7}$$

For the perfect information scenario in which the true demand is known in advance, PNLP reduces to a deterministic linear program. The RM system of our partner company uses a PNLP-based optimization module.

Overall, the simulation uses a warm-up phase of 120 days so as to reduce the impact of the initial state of the system on the revenue performance of our unconstraining techniques since they rely on availability of sufficient historic data. Within that warm-up period the booking system uses the $FCFS$ policy to admit bookings until the capacity limit is reached.