Revenue management strategies and Booking.com ghost rates: a statistical analysis

Strategie di revenue management e Booking.com ghost rates: un'analisi statistica

Cinzia Carota, Consuelo R. Nava, Marco Alderighi

Abstract To investigate hotel revenue management (RM) intensity, a dedicated database is constructed from Booking.com. A critical issue in the crawled hotel room rates is the presence of missing values for certain types of rooms, weeks of stay and booking days. Such unobserved rates are termed "ghost rates", since they may result from RM strategies and not only from room unavailability. Our goal is to reconstruct ghost rates. To avoid bias induced by deterministic imputations, we adopt a stochastic approach to multiple imputation that exploits the time-series cross-section structure of the sampled rates and domain-specific prior knowledge, thereby improving the plausibility of imputed values and preserving, at the same time, the statistical properties of the completed data. Then, we propose a clustering of room types, based on the completed rates, useful to study RM strategies at hotel level.

Abstract *Per indagare l'intensità del revenue management (RM) nelle strutture alberghiere italiane si è costruito un database da Booking.com. Il principale problema nell'analisi delle tariffe delle camere degli hotel campionati in questo modo è la presenza di valori mancanti, per determinati tipi di camera, settimane di soggiorno e giorni di prenotazione. Tali prezzi non osservati sono detti "ghost rates", dato che potrebbero essere il risultato di una strategia di RM e non derivare semplicemente dalla mancata disponibilità di una camera. Il nostro obiettivo è ricostruire i ghost rates. Al fine di ovviare alla distorsione indotta da tariffe imputate attraverso metodi deterministici, proponiamo un approccio probabilistico all'imputazione multipla*

Cinzia Carota
Università degli Studi di Torino, Dipartimento di Economia e Statistica, Torino (Italy) e-mail: cinzia.carota@unito.it

Consuelo R. Nava
Università della Valle d'Aosta, Dipartimento di Economia e Scienze Politiche, Aosta (Italy) e-mail: c.nava@univda.it

Marco Alderighi
Università della Valle d'Aosta, Dipartimento di Economia e Scienze Politiche, Aosta (Italy) e-mail: m.alderighi@univda.it

che sfrutta la struttura time-series cross-section dei dati e informazioni a priori specifiche di area, migliorando in tal modo la verosimiglianza delle tariffe imputate e salvaguardando, al tempo stesso, le proprietà statistiche dei dati completati. Successivamente proponiamo un clustering degli andamenti delle tariffe completate per i diversi tipi di stanza utile a studiare le strategie di RM a livello di hotel.

Key words: ghost rates; revenue management; stochastic multiple imputation; time series clustering

1 Introduction

Developed especially in airlines, hotels and rental car industries, the revenue management (RM) is a set of tools and pricing strategies designed to allocate the right capacity, to the right customer, at the right price, at the right time [7]. RM strategies are employed to take advantage from customer heterogeneity. In order to better price discriminate and, therefore, extract rent from consumers, product differentiation is often undertaken. In this frame, RM strategies also concern the use of dynamic pricing to maximize revenues [5, 10].

Only recently, hotel RM received a considerable attention, also because of the diffusion of online booking platforms such as Booking.com. Indeed, the Internet deeply changed the ways how hotels communicate and fix their room rates or room availabilities [6]. As in the case of airlines, hotel products are perishable, the room demand varies over time, and hotels have, at least in a short term, high fixed costs and low variable costs. During periods of high demand, as a results of RM strategies, rooms are usually affordable only to customers with higher willingness-to-pay, while during periods of low demand, room rates become lower.

In this paper, we aim at investigating the hotel RM activity through the study of their room rates as a function of the week of stay and the number of days between the booking and the check-in time, shortly referred to as day left. To this aim, exploiting a dedicated webcrawling system, we collected Italian hotel rates from Booking.com. Three, four and five star hotels are randomly sampled among the ones located in 22 pre-selected Italian touristic cities [4]. However, during the inspection period, missing room rates are observed. Given their numerousness, which opens challenging issues, and in order to prevent rate variability bias, we propose a method to pre-process data intended to impute missing rates.

Indeed, hotel missing rates result from two orders of reasons. On the one hand, all rooms of a given type are already occupied. On the other hand, manager decided to add or to take-off room types from the booking platform based on different motivations. First, since the commissions on Booking.com are really high, RM analysts try to maximize hotel visibility and simultaneously minimizing the sales. Second, RM analysts would prefer to sell rooms using alternative channels and only close to the date of stay they add an extra channel by putting last-minute offers on Booking.com platform. Finally, in order to offuscate the hotel room availability,

they choose to sell only a limited number of rooms on the platform and, therefore, once sold and before replaced, some room types are unavailable online. Thus, the missingness represents it-self a variety of RM strategies. Here all these missing values are termed "ghost rates" (GRs) to capture the idea that the reasons behind such unavailable room rates are unknown, but crucial to understand hotel RM tool mix and RM intensity [1].

Our major contribution is a statistical formulation of the RM analysis in hotel industry. Given that RM tools include techniques devoted to control room availability as opposed to techniques directly acting on room prices, we start our analysis by exploring the missingness map in the sample of rates crawled from Booking.com as detailed at the beginning of the next section. Preliminarily, we distinguish patterns of missingness whose generating mechanism turns out to be not at random (MNAR) from patterns whose ghost rates belong to a class of missingness ignorable from a statistical point of view.[1] For MNAR ghost rates we propose a series of ad hoc statistical models (see, e.g., [8]), also suggesting improvements in the data collection process, useful to quantify a series of specific non-pricing RM tools. For the remaining ghost rates, required for an unbiased and efficient analysis of most pricing tools, we adapt a well-established multiple imputation program (Amelia II) so as to introduce domain-specific knowledge through suitable prior distributions, in addition to smooth time trends, shifts across cross-sectional units, and correlations over time and space. Our multiple imputation model is much more flexible than the one in [1], while we consider the same ANOVA model to analyse the variability of the completed rates. Since, in general, hotel managers define specific pricing strategies for different room types, some of which with a higher dynamics in rates than others, we also suggest an appropriate clustering of hotel room types, relying on the completed data, useful to study RM at hotel level.

2 Data collection, ghost rates imputation and room types clusterization

The data collection process started the 1st of May 2018 and ended the 10th of August 2018. This webcrawling system generates a database composed of 1100 Italian hotels, for a seven-day booking period from the beginning of July 2018 to the end of October 2018. Room rates have a multiple index, h, t, d, w, with h denoting the hotel, t the room type, d the number of days left, and w the week of stay. The set of available room types in hotel h, in a given week and day left, is $C_{h,d,w} \subseteq C_h$, where C_h is the set of all room types in that hotel. $C_{h,d,w}$ is dynamically adjusted for each d and w. For each hotel, RM allocation strategies induce a three-dimensional room rate matrix, $\boldsymbol{R}$, with generic element $r_{h,t,d,w}$ and potential GRs. Indeed, our sampled $\boldsymbol{R}$ exhibits 169,973 GRs (30.83%), ranging from a minimum of 0% to a maximum of

[1] Missing completely at random, MCAR, or missing at random, MAR, by applying the Little's test and by visual inspection, respectively. The MAR assumption is also indirectly checked by using simulated missing data (see, e.g., right panels in Figure 2).

81.86% GRs per hotel. Figure 1 illustrates the great variability of room type rates in different hotels for the minimum observed day left, in the absence (hotels A and B) and in the presence (hotels C and D) of missing values with very different patterns of missingness. Ghost rates in both hotels C and D are classified as MAR, with missing depending on the week of stay (rooms booked before May 2018).

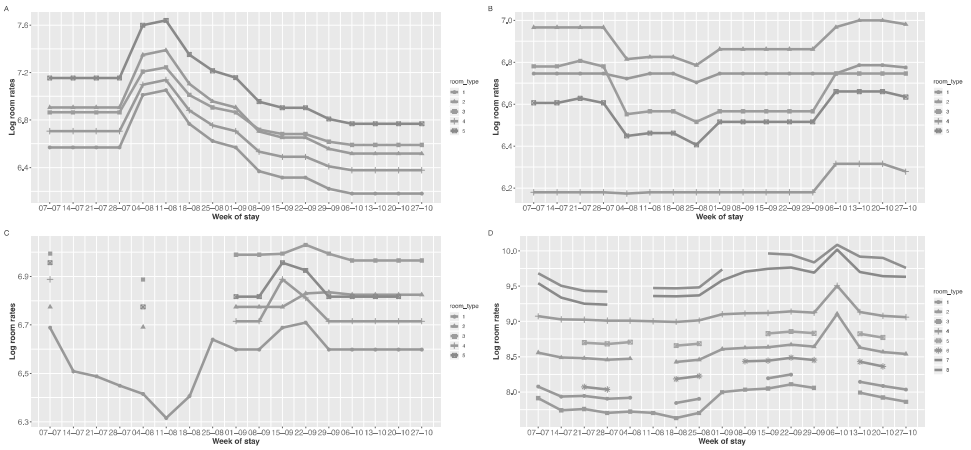

Fig. 1 Room type log-rates in four different hotels in the presence of the minimum observed day left. Hotel A is a four star hotel located in Ischia (seaside), hotel B is a four stars hotel located in Naples (art city), hotel C is a four stars hotel in Turin (art city) and hotel D is a five stars hotel located in Rome (art city).

.

Data imputation for multivariate time series can be a challenging problem, especially when temporal patterns as well as missingness patterns are quite different. The crude frequentist, hotel-specific approach [1] imputes GRs relying on 1100 OLS estimates:

$$\log(r_{h,t,d,w}) = \alpha_{h,1} + \boldsymbol{\beta}_h \boldsymbol{x}_h + \varepsilon_{h,t,d,w} \tag{1}$$

with $h = 1,\ldots,H = 1100$; $t = 1,\ldots,T_h$; $d = 1,\ldots,D$; $w = 1,\ldots,W$. The dependent variable is the natural logarithm of $r_{h,t,d,w}$, while the covariates $\boldsymbol{x}_h$ are the room type id codes represented by a set of $T_h - 1$ dummies whose effects, $\boldsymbol{\beta}_h$, are additional intercepts with respect to the reference one, $\alpha_{h,1}$. Finally, $\varepsilon_{h,t,d,w}$ denotes a Gaussian white noice. Here, we enrich such imputation model in various ways, in order to take advantages of both information embedded in the entire dataset and domain-specific information. Our more flexible model,

$$\log(r_{h,t,d,w}) = \sum_{i=0}^{K} \beta_{h,t,i} w^i + \gamma_{h,t} d_{h,t} + \delta_{h,t} b_{h,t} + \zeta_{h,t} s_{h,t} + \eta_{h,t} f_{h,t} + L^- + L^+ + \varepsilon_{h,t,d,w}, \tag{2}$$

reduces to eq. (1) for $i = \gamma_{h,t} = \delta_{h,t} = \zeta_{h,t} = \eta_{h,t} = L^+ = L^- = 0$, $(\beta_{h,t,0} = \alpha_{h,1} + \beta_{h,t})$. For each hotel h, eq. (2) considers the multivariate, weekly-spaced, time series of log room rates, whose dimension T_h accounts for the so-called *second-degree price discrimination* while the K-th degree polynomial of the time index w is introduced to account for *peak-load pricing*. Further variables considered in the imputation model are the day left, d, to capture *inter-temporal price discrimination*, the maximum number b of guests in the room, the room size s, the free cancellation option f, together with lags and leads, L^- and L^+ respectively, of the observed log-room rate.

In addition, since hotel room minimum and maximum observed rates concur to form customer *reference prices* of a specific hotel room type, such information is suitably embedded in a prior distribution, thereby considering a Bayesian version of model (2). Finally, instability of the EM imputation algorithm is avoided by slightly shrinking the covariances among the variables toward zero by means of a so-called ridge prior .

The quality of imputations generated by our Bayesian model is explored in left panels of Figure 2, while we provide a twofold indirect check of model adequacy (and plausibility of all underlying assumptions) in right panels. There, focussing on hotel B, we show imputations of ghost rates simulated by deleting observed values so as to exactly replicate in hotel B some patterns of missingness observed in hotel D (top right panel, B_1) and by deleting completely at random 20% of the observed rates (bottom right panel, B_2). When comparing imputed rates (connected by thinner lines) with the deleted true ones (imported from Figure 1, panel B), despite the low degree of the polynomial function (K=1), we observe quite good imputation results in both B_1 and B_2 cases.

We then apply to the completed time series of rates a clustering of room types governed by a dissimilarity measure able to capture and compare the higher-level dynamic structures describing the global behaviour of the series. In particular, to group homogeneous rate patterns useful to study RM activity at hotel level, we realize a hierarchical clustering based on the Ward's method and a distance constructed by considering the partial autocorrelation functions with geometric weights decaying with the lag [9]. The selection of this distance ensures the reasonable partition of hotel rooms presented in Figure 3. The goal of this clustering is to group rooms with affine rate patterns, not necessarily associated with similar rate levels. Indeed, different and sometime counterintuitive clusters are obtained when considering distances based on raw data, on correlation or on discrete wavelet transform.

References

1. Alderighi, M., Calabrese, M., Christille, J.M., Nava, C.R., Salvemini, C.: Room Rates and Hotel Price Fairness, mimeo (2019)
2. Bauer, J., Angelini, O., Denev, A.: Imputation of Multivariate Time Series Data - Performance Benchmarks for Multiple Imputation and Spectral Techniques (July 3, 2017). Available at SSRN: https://ssrn.com/abstract=2996611 or http://dx.doi.org/10.2139/ssrn.2996611

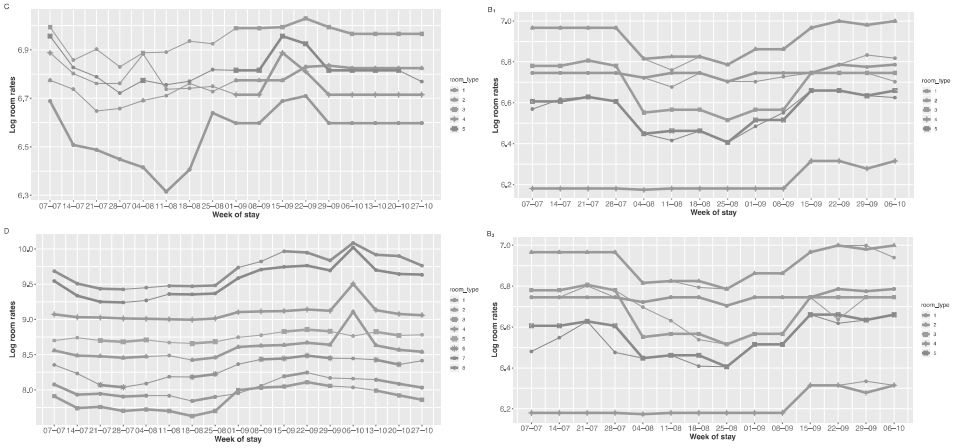

Fig. 2 GR imputations according to the Bayesian version of model (2) in hotels C and D (left panels) and in hotel B (right panels), in the presence of GRs generated so as to replicate in that hotel five patterns of missingness observed in hotel D (panel B_1) or completely at random (panel B_2). In all panels, imputed rates are connected to the observed ones by thinner lines.

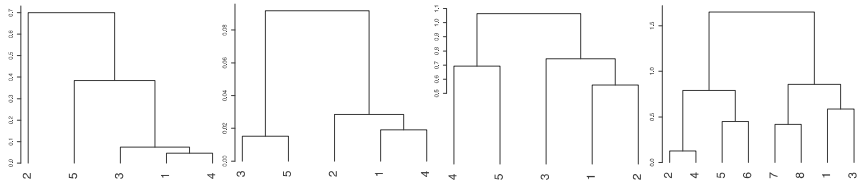

Fig. 3 Dendrogram obtained from a hierarchical clustering of completed room rates multiple imputation in hotels A, B, C and D.

3. Honaker, J., King, G., Blackwell, M. (2011). Amelia II: A program for missing data. Journal of statistical software, 45(7), 1-47
4. ISTAT: Movimento turistico in Italia. ISTAT, Statistiche - report, 1-22 (2017)
5. Ivanov, S., Zhechev, V.: Hotel revenue management: from theory to practice. Varna: Zangador (2014)
6. Kannan, P. K. K.: Dynamic pricing on the Internet: Importance and implications for consumer behavior. International Journal of Electronic Commerce, **5**(3), 63-83 (2001)
7. Kimes, S. E. and Wirtz, J.: Has revenue management become acceptable? Findings from an international study on the perceived fairness of rate fences. Journal of Service Research, **6**(2), 125-135 (2003)
8. Molenberghs, G., Verbeke, G.: Models for discrete longitudinal data, New York: Springer-Verlag (2005)
9. Montero, P., Vilar, J. A.: TSclust: An R package for time series clustering. Journal of Statistical Software, **62**(1), 1-43 (2014)
10. Talluri, K. T., van Ryzin, G. J.: The theory and practice of revenue management. International series in operations research & management science. Kluwer Academic Publishers, Boston, MA, 2-14 (2004)