FIELD EXPERIMENTS (AND LACK OF THEREOF) FOR TESTING REVENUE STRATEGIES IN THE HOSPITALITY INDUSTRY

David López Mateos*
Chief Science Officer
Pace Revenue Management
London, UK
david@paceup.com

Maxime C. Cohen
Desautels Faculty of Management
McGill University
Montreal, Canada
maxime.cohen@mcgill.ca

Nancy Pyron
Independent Consulting
nancy.pyron@gmail.com

August 1, 2020

ABSTRACT

This paper outlines opportunities that field experimentation can bring to accommodation managers. It also describes specific types of experiment designs that can help exploit those opportunities and increase the adoption of field experimentation. Field experimentation (often through randomized testing) has been widely adopted as an optimization technique in product design and marketing in several industries. It is also considered as the golden standard for causal inference and thus a critical tool for decision makers. As such, companies have successfully used field experimentation to reduce costs, increase revenues, and maintain an edge in their customer experience in highly competitive environments [1]. However, a number of optimization problems with a rich academic and commercial history have resisted this push. In certain industries, such as hospitality, to the authors' knowledge, there is little publicly documented work detailing results of field experiments applied to revenue management, and the use of such tools remains the privilege of big corporate brands with a small overall market share. This happens in a $500 billion industry in which vendors and academics alike claim that price optimization can yield uplifts of 10% in revenue [2]. This paper discusses the likely causes of the sparse adoption of field experimentation for revenue management in hospitality. By explicitly addressing the complexities of revenue management, and outlining specific experimental designs aimed at handling those complexities, this paper aims to start a public conversation about experimentation in hospitality that should benefit the industry as a whole.

1 Introduction

Statistical field experimentation, also referred to as controlled experimentation, control/treatment testing, A/B testing, and split testing, has been widely adopted as a development tool in product design for web-based software consumer products across industries [3]. As the golden standard for causal inference, it also plays a key role in strategic business decisions, such as choosing the right algorithm to route data efficiently [4] and optimizing the redesign of a website [5]. Generally speaking, causal inference and optimization through experimentation have been more prevalent in purely experimentally-driven areas in which no alternative theory exists (e.g., product design). Its adoption is limited across products that feed from strong academic fields with a rich history. However, even in those cases, the academic work and resulting models may have a hard time accurately representing commercial settings (e.g., in the case of pricing, goods can be highly differentiable so that analytical models might fail at faithfully representing realistic scenarios).

*Corresponding author

Revenue management is a set of optimization techniques and heuristics developed originally in the 1970s and 1980s to help increase revenues (and ultimately profits) through appropriate allocation and pricing of seats in airlines [6]. Since then, it has evolved to adapt to a multitude of complex commercial needs (product bundling and unbundling, personalized promotions, lifetime customer value estimates, etc.). The techniques and models have been applied across the perishable industry spectrum, including hospitality [7, 8], rental cars [9], trucking, rail cargo, and airline cargo with various degrees of market penetration.

As adoption and complexity have increased, so have alternative solutions to the original optimization problem, whether considering more realistic formulations of the problem [10] or using modern techniques in decision making [11]. In this evolving landscape, however, there has been limited focus in the literature on evaluating the merits of different implemented approaches in realistic settings, with some exceptions [12, 13, 14]. In the hospitality industry, the authors could not find any public article describing in detail a field experiment to decide amongst several possible revenue strategies.[2] Instead, one can regularly find mentions of verification procedures and observed uplifts of up to 10% in business conferences and marketing materials.[3]

One of the goals of this paper is to provide a starting point for revenue managers and data scientists who are interested in using field experimentation to increase revenues. To do so, this paper first outlines some of the complexities of revenue management in hospitality that hinder the widespread application of field experimentation. It then identifies several opportunities that field experimentation can offer to revenue managers. Finally, it proposes three concrete experimental designs that can be used to successfully exploit those opportunities, and it highlights relevant empirical methods that can be used to extract valuable information from field experimentation.

Specifically, in Sec. 2, we discuss the current state of online field experimentation, first in the pioneering big technology firms and then focusing on hospitality. We report a brief literature review as well as undocumented—but generally known—practices. In Sec. 3, we describe hotel operations in the context of pricing and revenue management, exposing some of the complexities which are inherent to the hospitality industry, and outlining common practices for measuring the quality of pricing decisions without using experimentation. In Sec. 4, we clarify the potential benefits that experimentation can bring to hotel revenue managers and other stakeholders. We also discuss some of the key challenges in this context, explaining the limited adoption of experimentation in hospitality. Finally, Sec. 5 presents three experimental designs that can help overcome those challenges, with the ultimate hope to stimulate the use of field experimentation in the hospitality industry.

2 The current state of online field experimentation

The rise of the Internet in the late 1990s and the enhanced data collection and analysis capabilities made it a fertile ground for the development of field experimentation. In software as a service applications, companies offer their product via hosting on the cloud, controlling at all times what version of the product is shown to which customers. Companies also have the capability to analyze the interactions of the user with each version of the product, hence an ideal setup to understand which version of the product is the most successful.

It is not surprising that, in this context, online field experiments have become ubiquitous amongst big technology firms as an indispensable tool for software development, algorithm optimization, and business decision-making. Companies such as Facebook, Google, Netflix, and Airbnb report that at any given time, thousands of field experiments are running in their products [15]. In certain mature areas, such as product design and marketing, tools that aim at commoditizing field experiments have started to appear, even though the extent of their usage by smaller companies remains unclear [16]. In Ref. [3], the authors report illuminating business examples of successful applications of field experiments. They also claim that one of the primary reasons of why online field experiments are such an important tool is the fact that: "We are poor at assessing the value of ideas." We next briefly discuss four common applications of field experimentation as well as report a more detailed overview of the current situation in the hospitality industry.

Product design Assessing the quality of a specific product versus another is not a simple task. Indeed, beauty is on the eye of the beholder and the same can be said about how appealing a product is perceived. In an online context, one can define several usage metrics, such as the frequency of logins and the time spent on a specific webpage. One can then aim to design a product that optimizes any set of such metrics. To formally compare the different products, field experiments can be a great tool. Airbnb, for instance, has described in great detail how the success of visual and feature changes to their search website is tested via online field experiments, including major redesigns as the one performed in

[2] Ref. [12] explains that a field experiment was used at IHG to measure the performance and provides a brief outline of the experiment, without reporting the details.

[3] See, for example, https://www.duettocloud.com/library/efficient-yielding-strategies-lead-to-big-gains-for-coast-hotels.

their search website in 2013 [5]. Similarly, Uber has been using a sophisticated experimentation platforms for several types of decisions [17].

User experience A user's experience is not fully determined by the product design. Sometimes, algorithms are developed to ensure that the experience is tailored to the customer's needs and preferences. Those algorithms are designed with a specific goal in mind, but understanding whether a modification has reached the desired outcome is not easy. Netflix, for instance, has used field experiments to decide which streaming quality to use or the type of artwork to display in order to entice users to watch content [4, 18]. Companies like ride-sharing are also routinely using field experiments to understand what is the best way to compensate customers who experience a low quality of service [19, 20].

Marketing As a business function driven by concrete metrics in most digital businesses, marketing is the perfect setting to apply online field experimentation. The interplay between marketing campaigns and other business units can make the results of an experiment hard to interpret (as we will discuss in Secs. 3 and 5). However, many customer relationship management (CRM) systems offer integrated tools to help with the set-up and analysis of field experiments. For simple marketing campaigns, Google offers a free service, called Google Optimize, which randomizes web traffic into your website and an alternate website. It then uses the data to calculate which of these two websites performs best along several relevant metrics. Marketing campaigns and customer nudges often rely on field experiments in various contexts including customer referrals [21], carpooling incentives [22], and philanthropic campaigns[23].

Pricing Finding the right price for goods sold online or even purely web-based products, such as streaming videos and video games, seems like an application where field experiments can quickly and unequivocally contribute to the bottom line by helping increase revenues. In fact, several businesses offer the capability to test product prices for small retailers selling on the Amazon marketplace.[4] In reality, laws and regulations in many countries often prevent discriminatory pricing, and thus make field experimentation in its simplest form not possible. However, as more businesses have adopted complex pricing algorithms that adjust prices based on price elasticity estimates, the question faced by pricing departments has often become: "What is the right pricing algorithm or strategy?" More sophisticated field experiments that stay far from discriminatory pricing have been successful at answering this question, for instance in the airline industry [14], in online retail [24, 25, 26], and even in physical retail stores [27, 28].

In most of the examples above, randomization is an important aspect of the experiment as it helps ensuring a fair experimental design that alleviates imbalances between the treated and control populations. Randomization is not possible in all settings, but that does not mean that field experiments cannot be used to establish causality and measure the effect of a treatment. When randomization is not possible, one often refers to field experiments as quasi-experiments, and the design of the experiment becomes particularly important. The use of quasi-experiments is less widespread among the big technology firms, due to their larger design overhead, but it is still being used frequently in companies like Netflix [29].

In summary, field experimentation has been used for years by big technology firms across several business functions. While some of the tools used for randomized testing and analysis have been made available to smaller companies, defining the relevant metrics and interpreting the results correctly requires domain expertise that is often available to large tech-savvy corporations. In situations where randomization is not possible, the experiment design becomes more important and it requires a higher level of expertise. This makes it even harder for smaller companies to add this fundamental tool to their decision making processes. Limited opportunities for randomization and a relatively small scale are two contributing factors that hinder the adoption of field experiments for pricing in the hospitality industry.

Experimentation in hospitality

In the context of the hospitality industry, it is important to differentiate between two types of businesses: (i) the ones who own (or operate) the inventory (i.e., hotel rooms), or *operators*, and (ii) the ones who do not and instead focus on helping operators, or *support businesses*. While this distinction is not perfectly accurate and may be disrupted in the future by new businesses [30], it is useful for the purpose of this paper.

Amongst support businesses, there are big technology firms such as Booking.com, Expedia, and Airbnb that provide primarily cloud-based software products. As it can be expected, adoption of online field experimentation amongst such companies is broad. As mentioned before, Airbnb is a good example of a technology firm that is using field experiments to improve their website. Interestingly, field experiments are even more pervasive in Airbnb's culture and have been described as a key element in their growth strategy [31]. Airbnb also provides pricing suggestions for their hosts [32].

[4] See, for example, https://www.splitly.com/amazon-pricing-wars/.

Several versions of the pricing algorithm may have been tested via online field experiments, even though limited details are available about these experiments [33].

Similarly, Booking.com has publicly advertised their experimentation culture in product design [34]. In fact, they have recently published a paper to discuss the way they enhance their experimental analyses to establish causality [35]. It is important to note that none of the above studies refers to pricing. Expedia has been less prolific about publicizing their experimentation efforts. In 2019, they described some work on multi-armed bandit optimization (a type of experimental optimization not far conceptually from randomized testing) [36]. The authors also know from private communications that Expedia has a small team of data scientists who focus on pricing algorithms (e.g., holiday packages and other product offerings for which they have some pricing control). This team is likely to use online field experiments for testing their algorithms, but there is no public information on this topic.

Amongst operators, the situation is quite different. First, their scale is smaller relative to the scale of support businesses. Specifically, the two largest hotel chains in the world by revenue (Marriott and Hilton) had a combined market capitalization of $81B at the end of 2019 and accounted for less than 10% of the rooms in properties with 20 or more rooms. In contrast, Booking.com and Expedia had a combined market capitalization of $102B and their revenues accounted for over 70% of all major online travel agents.[5] In addition, operators are not technology businesses at their core. The combination of these two structural differences implies that most operators are ill-equipped to introduce experimentation as part of their business processes, and the few that have the right size do not publicize their methodologies and results.

It does not mean that field experiments have not been used by operators. With many sales and marketing campaigns happening online, digital marketing products that target the hotel industry started offering randomization and significance analysis as part of their solutions. However, there is no public figures on the adoption levels of such tools in hotel marketing. Anecdotal evidence gathered by the authors suggests that this type of tools are primarily used by hotel chains that manage more than 200 rooms and have a sizable marketing department with data analysts. This includes all major hotel chains, but also several small to medium-sized chains managing more than 10 properties.

In addition, to the authors' knowledge, it is common for hotels to run field experiments as part of major revenue management systems upgrades (as reported in Ref. [12]). In this case, experimentation as part of the decision-making process is not a deep cultural characteristic as in the support businesses, but this conveys a desire amongst operators' business leaders to be part of this technological shift. Despite this desire, there is no commercial system for hotels to help design and run field experiments in order to guide strategic pricing decisions.

3 Revenue strategies and operations in hotels

Field experiment design requires a detailed understanding of how goods are sold and what type of questions the experiment aims to address. Hotels have a large market share in hospitality, a complex product distribution, an evolving technology infrastructure (or stack), and a diverse set of approaches to optimize their product offerings. This section cannot do justice to all these topics. Instead, we attempt to introduce succinctly some of the key concepts as a way to prepare the discussion in Sec. 4 on opportunities and challenges faced by field experiments for revenue strategies. More in-depth discussions on each of the topics introduced in this section are beyond the scope of this paper. The first part of this section focuses on describing common concepts in the practice of revenue management for hotels. It is then followed by a description of several prevalent features of the technology stack in most hotels, leading to a discussion on hotel pricing operations: what systems and people are involved and what type of decisions they commonly face? The section concludes with a discussion of KPIs that are often used to evaluate the quality and impact of pricing decisions.

3.1 Background on hotel revenue management

The practice of optimizing inventory allocation and pricing in hotels was originally inspired from the airline industry. The theory of inventory allocation or *inventory management* was developed in the 1980s, when airlines would commit to certain fixed-price fares, and their only control was whether or not to make these fares available. In this context, hotels used demand forecasts for each price point to determine how many reservations should be accepted at each price in order to maximize revenue. This is illustrated in a simplified example in Fig. 1. On the left panel, we can see how an inventory management calculation accepts many bookings at price point 1 (p_1). Once all those bookings have been confirmed, it stops offering rooms at p_1 and accepts a few bookings at price point 2 (p_2). In the situation illustrated on the right panel, the demand forecast is such that a larger number of customers are willing to pay a higher price, and thus fewer bookings are accepted at p_1, allowing 50% of the inventory to be sold to customers who are willing to pay

[5] See, for example, https://medium.com/traveltechmedia/the-state-of-online-travel-agencies-2019-8b188e8661ac.

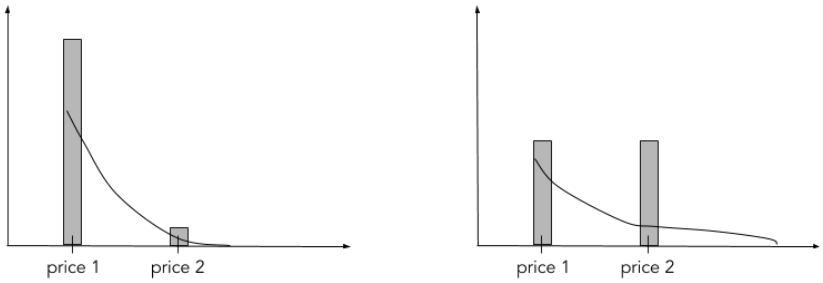

Figure 1: Illustration of how inventory controls try to optimize demand for two different willingness to pay distributions. The x-axis represents price and the y-axis represents how many people are willing to pay at most that price for the product (hotel room.) The hotel room can only be offered at two different prices, price 1 and price 2. The bars represent the number of bookings accepted by the inventory management algorithm at each price. On the left panel, the willingness to pay distribution (solid line) falls steeply, so that there is a low demand at price 2. Most of the available inventory is sold at price 1 before price 2 is even made available. On the right panel, the willingness to pay falls less steeply, so that inventory controls accept less bookings at price 1 to allow more customers to pay price 2.

more. The practice of controlling availability of rooms at different price points is also called *yield management* or even *revenue management*.

We note that this optimization process uses a binned estimate (the forecasts) of the underlying willingness to pay distributions (shown as solid lines in Fig. 1). Of course, if one can vary the prices (i.e., the location of the bins), the optimization problem changes and it can find a new optimal solution. This is illustrated in Fig. 2. In the airline industry,

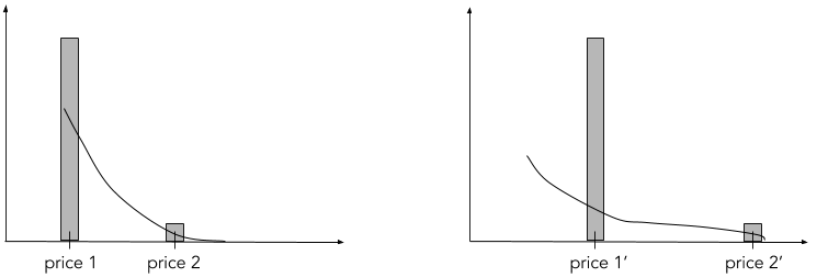

Figure 2: As in Fig. 1, the hotel room can only be offered at two different prices, but these prices can be decided based on estimates of the willingness to pay distribution. On the left panel, the willingness to pay distribution (solid line) falls steeply, so that the two prices are close together, one probing the core of the distribution, the other probing the tail. Most of the available inventory is sold at price 1. On the right panel, the willingness to pay falls less steeply, so that pricing controls increase both prices and separate them further apart.

the process of optimizing the price points at which each fare is offered is often called *pricing management* and is done by *pricing* tools. Pricing tools for hotels originally served a similar purpose, for example, by shifting the price points according to external data sources (e.g., competitors' prices). However, their most important role was to help hotels cater both leisure customers (that could accept dynamic prices) and business customers (that had been promised a fixed price). In hospitality terms, pricing became an important part of revenue management when the need arose to manage simultaneously dynamic and static *rates*, as we elaborate in the next paragraph.

It may sound counterintuitive, but at any given time, the same hotel room can be sold at more than one price. A *rate* is a set of prices associated to a specific set of potential customers. For example, the public rate might include all the prices at which a room has been sold publicly either through the hotel website or via online travel agencies (OTAs) such as Booking.com or Expedia. Public rates can be updated at any time, and hence are dynamic. Big hotels also often negotiate a private rate with corporate clients who commit to a certain volume of bookings in exchange of a fixed price.

These negotiations do not happen regularly (e.g., once a year), and hence these rates are static. We elaborate more on hotel rates in Sec. 3.3, and a more comprehensive discussion can be found in Ref. [37]. The key takeaway from the above brief introduction on hotel rates is that some part of the demand can be controlled through pricing, whereas some other part can only be controlled through inventory management. A common way for hotel revenue management systems to deal with merging inventory and pricing controls is by creating *hurdles* (also referred to as bid-price controls in the airline industry [38]). Hurdles are output of the inventory management system and may require that price point 1 is not available. Simultaneously, the pricing system might conclude that rooms should be sold at price $1'$, using the notation of Fig. 2. Public rates will then be sold at price $1'$. The hurdle will effectively close the availability of all rates sold at price 1 or lower. This will include static rates as well as other private rates which are derived from the public rate (in this example, any rate sold at a discount higher than the difference between price $1'$ and price 1).

The problem of optimally pricing hotel rooms is further complicated by the fact that demand is often segmented in other dimensions. In previous paragraphs, we have considered one *stay night*, or the most granular product offering for a hotel (i.e., one room for one night). Naturally, the same room sold on a different stay night can be seen as a different product and will experience a different demand pattern, even though the demand time series will likely show a high level of auto-correlation. This specific room and many others (which are indistinguishable from one another for the potential guest) are typically on sale for a long time (e.g., one year). During that *booking window* the willingness to pay will also vary, sometimes in predictable ways, so that revenue management algorithms can try to anticipate this variation. An additional time dimension that is relevant for hotel revenue management is the *length of stay*. Customers typically do not reserve only one stay night, but several stay nights in a row. The type of traveller making a 2-day booking might be fundamentally different to the one making a 7-day booking, so a revenue optimization algorithm may want to account for this type of segmentation. Finally, a hotel will often split their product offerings in different room types (e.g., doubles, twins, suites). A practical revenue management algorithm will then need to consider that demand for these different room types might show some degree of cannibalization (assuming that the different room types act as substitute products).

The previous paragraph should make it clear that practical revenue management is far from the idyllic examples of Figs. 1 and 2. To handle complex demand patterns, diverse data sources with various degrees of accuracy, and other intricate operations that we will discuss in Sec. 3.3, hotels that price dynamically their rooms often use a revenue management system (RMS) to aid or even fully automate inventory and pricing controls. Those systems may use binned estimates of the willingness to pay distribution, or simple rules to follow competitors' prices. They may also include more sophisticated un-binned estimates of a multidimensional willingness to pay distribution over all the relevant time dimensions, or follow a model-free reinforcement learning approach to learn the optimal pricing and inventory management policy. The underlying implementation of these systems is beyond the scope of this paper. The crucial takeaway from the previous discussion is that people and software systems will make decisions about when to make static rates available and how to price dynamic rates. In the rest of this paper, we will refer to this combined set of decisions as the *revenue strategy* and we will endeavor to provide methods to assess the quality and impact of different revenue strategies. The next section provides a short overview of the software systems involved in implementing revenue strategies, whereas the following section discusses the key decision makers who decide those strategies and the types of decisions they face.

3.2 The hotel revenue management technology stack

The hotel technology stack has evolved over time and, with different hotels being at different stages of that evolution, it is hard to provide a one-fits-all description. That said, there are certain general concepts that apply to the vast majority of hotels.

A hotel needs an ultimate source of truth for all their data: reservations, prices, space configuration, restaurant expenditures, etc. This is, at its core, a data warehousing system,[6] but naturally many products can be built around that system. Today, a property management system (PMS) is often this data warehousing system, even though large hotel brands often use a central reservation system (CRS) instead, as a way of managing and aggregating a variety of PMS technologies across their hotels. For the purpose of this paper, such details are not important, and hence we will simplify by referring to the central data warehouse as the CDW. For more details on the intricacies of the hotel technology ecosystem, we refer the reader to Ref. [39]. The CDW will often have information about individual customers: previous stays, previous expenses, loyalty programs, etc. This type of information cannot be shared with external vendors to comply with the general data protection regulation (GDPR) and is also not often available before a booking is made. For big brands with loyalty programs, this information can be interesting in designing revenue strategies that are targeted to specific customers, but the focus of this paper is more on general strategies, so we will not explicitly refer to this set

[6] In practice, this is not always the case, and the information is spread between several systems (e.g., the PMS and the CRS), but for the purpose of this paper we can rely on this simplification.

of data in future sections. PMSs and CRSs also provide tools to serve functions that can be built on top of the data warehouse (such as front-desk, guest management or housekeeping tools). Alternatively, they integrate with other vendors that specialize on providing services to hotels, granting them access to a subset of this data.

The CDW stores prices (mostly for rooms, but could also store other prices such as restaurant and spa items). The CDW can be connected to OTAs and to the hotel's own booking engine to guarantee that the right prices are used for any incoming booking. This connection can be facilitated by other applications, such as the CRS or channel managers (CMs). OTAs have only access to a small fraction of the CDW data, mainly prices and room availability. Prices are stored and can be modified manually directly in the CDW. Modifications in the CDW may take some time to propagate to all the relevant systems. The exact amount of time varies between seconds and hours and is specific to the system. Most PMSs and CRSs do not provide automated pricing algorithms. As a result, most sizable hotels employ an RMS designed to ease the implementation of revenue strategies, either through a dedicated user interface with demand estimators, or by using more sophisticated algorithms that automate most pricing and rate availability decisions. Several hotels perform most of their price changes directly in the RMS. These changes propagate to the CDW and then to all the distribution channels. Hotels without a CDW (e.g., small hotels that manage bookings manually) will make most revenue strategy updates through their CM. Finally, some revenue departments of hotel brands will make revenue strategy updates through the CRS (which will get propagated to the PMS).

3.3 Revenue strategy decisions

In Sec. 3.1, we introduced two types of rates, static and dynamic. The reality of hotel operations is more complex with rates existing for every type of product and every type of customer. It is common for large hotels to have tens, or even more than one hundred, different rates. For example, two common public rates in the hotel industry are the refundable and non-refundable rates. A strategic decision might be whether to offer both rates and what should be the difference between these two rates. We can think of these two rates as one public rate that sometimes is sold with a cancellation insurance (refundable) and sometimes without. The strategic decision thus reduces to deciding the value of the cancellation insurance (which would be arbitrarily large if the refundable rate is not offered and zero otherwise) and the price of the public rate. A similar construction can be applied to rates with or without breakfast as well other add-ons.

In Sec. 3.1, we also discussed the example of a fixed-price contract with a corporate client, which is a static rate. In fact, every corporate client is likely to have a different rate. This rate is sometimes linked to the public rate (for example, offering a fixed percentage discount over the public rate). A strategic decision in this context is to decide which of these rates (if any) to make available.[7] It is also common for hotels to have negotiated rates for groups with a certain volume. These rates differ from negotiated corporate rates in that the negotiation can happen at any time, and apply to specific dates and rooms. Therefore, pricing decisions for these rates are made at the time that the group expresses interest in booking with the hotel. Finally, discount rates may be created by the marketing department to stimulate business through promotions. These rates are limited in time and could be either publicly available (everybody reserving during a specific time window will receive it) or private (e.g., customers who have previously stayed in the hotel receive an email with the promotional code).

We highlight that not all the rates are available in all channels. For example, to unlock a corporate negotiated rate, one may not be able to use an OTA, and instead will need to use the hotel website or call the hotel. Most contracts with OTAs have a rate parity clause [40], by which a hotel cannot publicly sell their rooms at a lower price through other channels (such as the hotel website), but sometimes hotels offer a lower rate through *opaque* channels (such as Priceline.com or Hotwire) that only reveal the hotel's name after it is booked. The rate parity clause can also be bypassed by OTAs working under an agency model (i.e., rather than operating as a marketplace, they can offer a certain volume at a certain rate, hence obtaining a level of pricing control) to directly sell the room, or sell it in a bundle with other products. The rate parity clause has been brought down by courts in some important tourist destinations, such as France.

Decisions related to the revenue strategy for each of these rates ultimately are the responsibility of a hotel management team. However, it is fair to say that the revenue management team controls the public rates and with the help of modern software will be reviewing and updating those rates multiple times a day.[8] While the revenue management team can decide at any time to make a negotiated or discount rate available or not, they will not always be able to control the price associated with these rates, which may have been negotiated by the sales or marketing teams. The interactions between rates, prices, decision makers, and distribution channels is illustrated with simplifications in Fig. 3. OTAs

[7] Contractually, hotels often cannot close the availability of such corporate rates unless the hotel is full, but they can reduce their demand by restricting bookings to a certain length-of-stay, which often preserves most of the demand at public rates.

[8] See, for example, https://www.paceup.com/blog/automation-launch/.

operating under the agency model are disconnected from any rates and decision makers because the price shown to the customer is not controlled by any of the rates or decision makers, even if the contract specifies that a fixed amount of money will be received by the hotel for each sale.

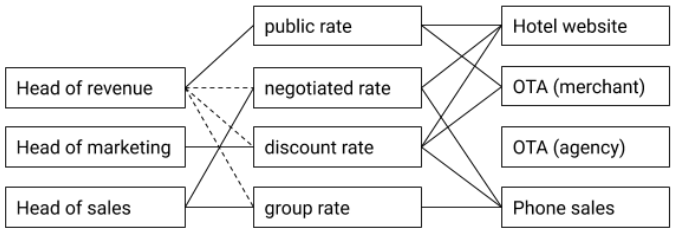

Figure 3: Simplified illustration of decision makers, main distribution channels, and their interactions with prices. A dotted line expresses that the decision maker can decide whether a specific rate is available, but not the price associated with that rate. The decision makers can be a single person (for a small hotel), three different people, or entire teams. They can be assisted by an RMS, and in some cases they will delegate most decisions to the system.

The complexity of these operations makes it challenging to design unbiased field experiments. For instance, because of rate parity, one cannot split traffic to the hotel website randomly and show two different prices to two different customers at the same time. If an experiment is running and it affects several nights, changes in negotiated or discount rate availability can prompt an unbiased design to become biased, as accepting a large group could significantly reduce the power of the experiment. For this reason, for a field experiment to be successful, it needs a deep level of pricing control (i.e., what prices to show, when, and where) and potentially an automated way to adapt to live changes (e.g., re-evaluate the expected power of the experiment and decide whether to extend its length, exclude specific problematic observations). No commercial product or system currently exists that is designed to efficiently run field experiments for hotels, but RMSs and potentially CRSs or PMSs are in a good position to become a central control center for experiments, given that they already have control over pricing.

These operations outline many of the strategic decisions faced regularly by revenue managers and the hotel management team:

- how to set the public rate of individual room types and nights for the next 6 to 18 months (often with most focus on the 2 to 6 future weeks),
- how to price cancellation insurance (refundable vs. non-refundable rate),
- how to price breakfast, high-speed Internet, resort fees[9] or other *ancillaries*,
- how to set the price for different groups,
- when to make negotiated corporate rates available,
- when to override the suggestions from the RMS (often to incorporate information that the RMS may not have, such as a local event that will impact demand),
- when to activate a promotion, who are the customers that should be offered this promotion, and how to reach out to them (e.g., email, website pop-up).

The above list is by no means exhaustive. In fact, revenue managers often find themselves making additional relevant decisions with or without the help of external systems, such as deciding which hotels are the relevant competitors and how to react to competitors' price changes. Overall, hotels may be routinely interested in testing several revenue strategies such as the ones discussed above. Field experiments are the golden standard for causal inference and can be a powerful tool to discern the effect of different competing strategies. Regardless whether the above strategic decisions will be automated in the future (which to a large extent we believe they will) or made by humans, hotel managers need ways to measure the quality and impact of these strategic decisions. Fortunately, the industry has developed standard processes and metrics to assess the quality of their revenue strategy.

[9]Some hotels will bundle several services, such as Internet, gym and spa access into a mandatory *resort fee, which is often undisclosed at the time of the booking.*

3.4 Common KPIs and measuring the success of a revenue strategy

As in most businesses, profits are the ultimate measure of success for a hotel. To compare properties of different size, it is common for hotels to normalize the performance metrics by the number of rooms in the property. GOPPAR represents the gross operating profit per available room.[10] This is a useful performance metric when considering the overall health of the business, but it is not of much use when trying to isolate the effect of decisions related to revenue strategy. If the operations are a lot costlier because of a new facility in the hotel, the GOPPAR may be lower either because of too low prices or because of the high cost of running the new facility.

Ideally, one would like to isolate the operating costs related to filling one room (variable costs per available room, or VCPAR), such as extra staff in the restaurant and cleaning costs, and calculate the room operating profit as the total revenue per available room (TRevPAR) minus the VCPAR. The reality is that isolating the operating costs related to filling one room from other operating costs is often not possible. Consequently, performance metrics for pricing decisions often ignore the variable costs. When comparing revenue strategies, this will often be valid, since all the strategies will have similar occupancy and variable costs. However, if one of the strategies involves an increase in variable costs (for instance, including in-room breakfast service), it could be quite relevant to include that cost when defining the KPI. Most of the time, however, this subtlety is ignored.

The revenue strategy decision at the room level often does not influence the total client expenditure in a hotel, unless the pricing decision considers past behavior of specific customers (or by using aggregated customer profiles). For this reason, RevPAR, or the revenue from the room sale normalized by the number of rooms is very often the preferred KPI to understand the impact of pricing decisions. It is worth noting that due to the GDPR (which prevents hotels from sharing their clients' personal identifiable information with software providers), private RMS providers cannot help improving TRevPAR by analyzing specific customer behavior and targeting customers who spend more, except with very coarse aggregation methods (such as the country of origin or the specific rate). Even for large hotel chains with internal RMSs, this type of optimization is only possible for the customers who are booking through loyalty programs, so TRevPAR remains an interesting KPI primarily for big leisure operations with a lot of client data and additional expenditures (e.g., casinos, cruises), and less relevant in measuring the performance of pricing strategies in the traditional hotel space.

RevPAR changes year-over-year are often used to compare the impact of a specific revenue strategy over time. However, without a sophisticated experimental design, changes in RevPAR can signal many trends, such as inflation, seasonality, or other macroeconomic factors. To decouple external factors from internal changes, hotels share their internal data with external companies (e.g., STR Inc.) that produce market indices to help understand the hotel's performance evolution with respect to a benchmark market. Such a benchmark market is typically a set of hotels in the same geographical region that offer similar amenities. The RGI (revenue generating index) normalizes the RevPAR by the aggregated RevPAR of the hotel's competitors. The RGI can be calculated daily or monthly, but most hotels use it as a monthly indicator. Although the RGI is not a perfect indicator of a change in revenue strategy, it is often used to measure the success of such a change (such as an upgrade of the RMS). Two similar indicators, the MPI (market penetration index), which measures the hotel occupancy (i.e., sold rooms divided by the number of available rooms) normalized by the benchmark market occupancy and the ARI (average rate index), which measures the ADR (average daily rate) normalized by the aggregated ADR of the hotel's competitors, are also monitored at the same time. This is not justified from the point of view of trying to maximize profits, but a large drop in the MPI would signal to management that their strategy (whether it is in pricing or not) might not be reaching potential clients [41]. Similarly, a large drop in the ARI might signal a change in how the hotel's value is perceived with respect to its competitors. All the KPIs discussed above are summarized in Table 1.

Despite the usefulness of these KPIs, it should be obvious that they lack granularity and that inferring causality from simple comparisons is at best a dangerous game. To our knowledge, small pricing decisions (such as when to activate a promotion or how to set the prices of ancillaries) are often left to the management team. Then, the quality of these decisions is typically measured by monitoring the above KPIs, without systematically accounting for confounding factors. For bigger decisions, such as the adoption of a new RMS, large teams may be involved in designing and running field experiments across both several properties and revenue managers before making a final decision. Smaller hotel groups do not have such a luxury, and opt for a wide range of ad hoc strategies, from testing two different RMSs in two different properties for a small time window, to being strongly swayed by other factors (such as selling price, ease of use, and endorsement by their revenue management team). In the next section, we discuss the opportunities that field experiments can offer in the context of hotel revenue management as well as some of the challenges that need to be overcome to exploit those opportunities.

[10]The word available in this context refers to all the rooms that can be sold, excluding the rooms that cannot be sold for a specific reason, such as being under renovation.

Table 1: Common KPIs to measure the effectiveness of revenue strategies.

KPI	Abbreviation	Description
Gross operating profit per available room	GOPPAR	Total sales revenue minus operating expenses normalized by the number of rooms that are available for sale
Total revenue per available room	TRevPAR	Total sales revenue normalized by the number of rooms that are available for sale
Occupancy	———	Fraction of available rooms that are sold
Revenue per available room	RevPAR	Revenue from accommodation sales normalized by the number of rooms that are available for sale
Average daily rate	ADR	Revenue from accommodation sales normalized by the number of rooms sold
Market penetration index	MPI	Hotel occupancy normalized by market occupancy
Revenue generating index	RGI	Hotel RevPAR normalized by market RevPAR
Average rate index	ARI	Hotel ADR normalized by market ADR

4 Opportunities and challenges for field experiments in hotel revenue strategies

In Sec. 3, we described the common revenue strategy decisions faced by a hotel manager as well as some of the relevant KPIs. In this section, we discuss several opportunities that field experiments can bring to hotel revenue management. We also report some of the practical challenges involved in experimental setups before proposing three concrete experimental designs that can help overcome some of these challenges in Sec. 5.

4.1 Opportunities

The effects of most of the decisions faced by a hotel manager (outlined in Sec. 3) have outcomes that can be measured through dedicated experiments. Field experiments can also help with other, longer-term strategic decisions, as well as decisions that might be relevant for RMS providers. This section discusses some of the areas of hotel revenue strategy that could benefit from field experiments, including estimates of the economic impact that these experiments could have for a hotel.

Ancillary pricing Should breakfast be sold at $5 or $6? For a hotel that sells 100 rooms at an average of $100 per stay night and has operating costs that make it break even at 50% occupancy, this is an impactful question. Such a hotel will earn a yearly profit of $1.095M when operating at 80% occupancy. If this $1 price increase has a negligible effect in people choosing to eat breakfast, it would then lead to a $29,200 profit increase (or 2.6%). The same question can be asked about cancellation insurance, high-speed Internet, phone charging cables, gym access, or any other product that the hotel may sell (often at a high margin) as part of its accommodation offering. A similar question can be asked about the value of room attributes (ocean view, bed size, etc.). In Ref. [42], the authors ran a field experiment to infer the economic value of such room attributes. Since the price sensitivity of such items is likely to vary slowly over time, field experiments are ideally suited to help optimizing their prices.[11]

Distribution channel optimization Should a hotel pay an extra 2% to an OTA to be listed in the first position when people search for hotels in the area? This is a common strategic question for hotels, which can easily receive 30% of their bookings through OTAs and pay them commissions of 10–20%. A 2% commission difference would actually correspond to a 2% difference in profits in the simple example above. The answer might differ for different hotels and different seasons, depending on how differentiated the hotel is, and how customers reach the hotel in their experience through the OTA. With the right bookkeeping and metrics, this question can be successfully answered with an experiment or a set of experiments that are run on a regular basis.

[11] Ancillaries are sometimes sold below cost, as a way of increasing conversion. If this is the case, it would require a different experimental setup, more akin to the one used to test retail promotions.

Promotion success Was a promotion successful at increasing the hotel's revenue? Generally, the goal of a promotion will be to stimulate the number of bookings, up to a level that is considered appropriate by the revenue management team. The promotion will be considered a success if it managed to obtain a specific target in terms of bookings. Ultimately, however, the objective of a promotion is to increase the overall profit of the hotel. Indeed, in our previous example, if a 20% promotion increases the occupancy by 20% during one low-season month, this translates into a 5% profit increase. It is thus important to understand whether 20% was necessary, or a 10% promotion would have been enough to trigger most of the extra bookings. Experimental setups are more challenging in this context, due to price parity clauses and the difficulty of comparing two different promotions on equal footing. However, with a careful design, for certain hotels this question can definitely be answered by using a field experiment, as we discuss in Sec. 5.

Choice of revenue management system Which software provider should a hotel use for their revenue management system? This is a strategic decision for many hotel managers are facing. Case studies from software providers often claim a 5–20% revenue increase, but "cherry picking" is likely to inflate those numbers (however, it is generally accepted in the industry that typical revenue management systems will on average increase revenues by 2–5%). Even without cherry picking, just because a revenue management system performs well on average, it does not necessarily mean that it will be successful for all hotels. In our previous example, a 5% revenue increase leads to a 13% increase in profits or $146,000. Typical software providers may charge $20,000–30,000 per year for their service for a property of this type, resulting in multi-million dollar annual contracts for a group of properties. The price might be an important factor in the decision of the hotel manager, but if the more expensive system increases revenues by 6% instead of 5%, the price difference becomes irrelevant. Running a field experiment during a short implementation period with the different systems can thus provide critical information in making this strategic decision. This discussion also applies to the situation where the hotel manager needs to decide whether to switch to a new revenue management system or not.

Price update frequency How often should a revenue management system update prices? RMSs update prices at different frequencies, from every 30 minutes to once a day, depending on the system. For a software provider working on one of these systems or a hotel chain with its own system, it would be interesting to determine whether the extra value delivered to customers is worth the investment in extra compute power and other overhead costs of designing a real-time system. It is hard to estimate the potential value of answering this question, but with RMS providers turning over hundreds of millions of dollars every year, it is clearly not only an academic question. Since software providers can often increase their update frequency at their will (with the praise of most customers), this is a question that can be easily answered by RMS providers using field experiments.

Acquiring external data and updating revenue management algorithms How should we adapt the current algorithms to increase the revenues? This is another question that is relevant for software providers with an RMS as part of their product offering, or for hotel chains with their own systems. A provider may consider, for example, using a new data source in their pricing algorithm to improve the accuracy of the price elasticity estimates, with the resulting recurring cost of obtaining those data and the cost of incorporating it in the current algorithm.[12] From a marketing perspective, claiming the use of a new data source might be interesting for a software provider, but if it does not generate more revenue for the hotels, it may not be worth the recurring cost. The provider can design a field experiment to decide whether the new data source adds value to the algorithm, and whether that value is worth the cost of acquiring the data. The results of this type of field studies could even be used to increase credibility during the sales process (e.g., when prospective customers are inquiring about the data used by the algorithm).

4.2 Data requirements

To fulfill the potential outlined in the previous section, an experimentation system would need to be enabled by a data processing and management system. The requirements of such a system are not straightforward, due to the complexity related to hotel operations (see Sec. 3). We discuss these requirements in this section.

An experimentation system, naturally, needs to access the data needed to compute the relevant metrics. Most often, one will be interested in measuring changes in revenue, so for every booking, the system should know how much net revenue was generated from that sale. Unfortunately, this is not always easy to calculate. For example, due to different practices in how OTAs charge commissions, the room revenue stored in the CDW for a given transaction may or may not include the commission fee (which is 10–20%). If the commission has not been subtracted from the sale price, the hotel will receive an invoice for the commissions due at the end of the month. Such an invoice may not be stored in the CDW, and instead is located in an accounting system. Even if this type of invoices are fed into the CDW, it is unlikely

[12] Historically, the value of a new dataset (e.g., competitors' rates, hotel reviews, social media data, macroeconomic indicators) would be evaluated according to the improvement in demand forecasts or price elasticity estimates, by using historical data. However, assigning a concrete economic value to the dataset is more appealing to the management team.

to be fully disaggregated (a lump sum commission will be subtracted, but one cannot know the exact portion of that commission that correspond to each sale), so the commission cannot be easily retrieved to calculate the net revenue for each transaction. Consequently, a change in the number of bookings from Booking.com or Expedia could affect the `RevPAR` without having any effect on the `GOPPAR`. In practice, the system will need the transactional room revenue stored in the CDW as well as any other quantities which are needed to compute the net room revenue (commissions, exchange rates, etc.) A *netting algorithm* can then help recover the net room revenue. Alternatively, the experimentation system would need to use an external system that implements a netting algorithm, such as certain CRSs.

Similarly, it is common for CDWs to include the revenue from the sale of ancillaries (e.g., breakfast) or the cancellation insurance (refundable rate) in the room revenue. To run experiments involving ancillaries, their price should be explicitly available, and the experimentation system needs the capability to change prices. Even if the experiments do not explicitly change the ancillaries prices, the system should be able to calculate the room-only revenue, and to guarantee that the ancillaries prices are kept constant throughout the experiment duration. In order to know the price of ancillaries and to be able to change them, the system needs to know the rate plan of each sale along with its specifics (i.e., what ancillaries they include), and to have the ability to modify the prices associated with those rate plans.

An experimentation system does not only need to calculate the performance metrics. It also needs to control for confounding factors that could seriously affect the results of the test. Knowing the meaning of each rate plan and being able to control over some of the different rate plans can alleviate this type of concerns. For instance, we can know the number of bookings for a given stay night that stem from a negotiated rate or make sure programmatically that no discount rate is available. The system will also need to keep track of manual overrides (e.g., on prices and room upgrades) and receive external inputs (e.g., which nights are affected by nearby events or holidays). While most of these are available in PMSs and CRSs, these requirements suggest that the experimentation system needs a high-fidelity two-way connection to the system acting as a CDW, as well as additional external data.

The list below summarizes the data requirements that an experimentation system for revenue strategy will need to access in most cases:

- transactional bookings, cancellations, and reservation updates with their associated revenues,
- distribution channel for each booking and the associated commission,
- rate plan of each booking and its meaning (e.g., corporate negotiated rate, group discount rate),
- manual changes to bookings (e.g., room upgrades) or prices (override of experimental setup),
- nights when demand patterns are expected to deviate significantly from their nominal value (e.g., holidays and special events like convention and large conferences).

In addition, the system will need to have the ability to change prices and to control the availability of rate plans. From a data perspective, an experimentation system could be built most naturally inside a PMS or a CRS. If the experiment is running on multiple properties, then the system needs to have reliable access to high-quality data across all the properties, which make the PMS less well-suited for such applications. Given the widespread use of RMSs that exert price control and sometimes even automate price and availability decisions, an experimentation system could also be built inside an RMS, with similar data requirements.

4.3 Practical challenges

Per the previous section, it should be clear that CRS, PMS, and RMS providers have all the information required to help hotel operators run online field experiments. If the potential economic value is as impactful as suggested in Sec. 4.1, then why are not all hotels avid experimenters? The answer is multi-faceted, involving various scientific, technological, and organizational challenges. The solutions to these challenges involve several stakeholders, as discussed below.

Scientific challenges

There is extensive literature on controlled experiment design and analysis. This follows from the fact that this topic is highly relevant to real-world applications. At the same time, each application needs to perform customized modifications to the experimental design to guarantee that the assumptions underlying the statistical analysis are satisfied. One of these key assumptions is the *Stable Unit Treatment Value Assumption* (SUTVA). Simply put, in the context of hotels, this assumption means that the *treatment* (i.e., the tested revenue strategy we are aiming to evaluate in the field experiment) should not affect the control units. Stochastic variations in performance due to unpredictable events (such as increased demand due to last-minute cancelled flights, or cancellations due to travel plan changes) do not violate the SUTVA, but seasonality, for instance, does (day-of-week dependence and other *reference effects*).

For example, we need to ensure that changing the revenue strategy of a specific room type in a given property will not affect the performance of rooms in the control group (those can either be different room types in the same property or the same room type in other properties). Specifically, if two hotels of the same brand with similar offerings are in the same geographical area, an improvement in one of the hotel's performance may be driven by a loss in the other hotel (i.e., *cannibalization*). As a result, what seems to be a positive (or negative) treatment effect might result in no change in performance if both hotels were treated. A similar type of effects that violate the SUTVA and are of special importance given the competitive landscape of hotels are *network effects*. Indeed, hotels, their competitors, and the potential clients can be considered as a network. In a network, what happens to one node (hotel) can impact the other nodes (e.g., competitors). In the extreme scenario where a competitor's strategy is to perfectly mimic the strategy of the treated hotel, the treatment effect might be totally washed out.

Fortunately, field experiments have been successfully run in networks and non-stationary environments in the past. In Sec. 5, we outline three experimental designs that can help overcome some of these challenges. We hope that presenting successful experimental designs can help increase the adoption of field experimentation in hotels. However, one of the conclusions from this discussion is that running a successful field experiment requires verifying that the underlying assumptions are met. We elaborate on how that can be achieved when discussing organizational challenges.

Technological challenges

At the time of writing, no commercial experimentation system that automates or assists the design, implementation, bookkeeping, and analysis of field experiments for hotel revenue management exists. Given the large overhead in data preparation and analysis that any data science organization faces, it is hard to see a wide adoption of experimentation practices without such a system. Certain experimental designs are better suited for manual setups, as we discuss in Sec. 5, but the burden will often be too high for small organizations.

Even if such a commercial system existed, it would have to be deeply integrated within a PMS, a CRS, or an RMS. However, the PMS and CRS industries are highly fragmented. Oracle-owned Opera (the largest PMS provider) serves around 10–20% of the hotels with 20 or more rooms,[13] but one can find thousands of PMSs that cover the remaining 80–90%. The CRS market is less fragmented, with three dominant products but still has tens of other competitors. In addition, most independent hotels, which make up around 50% of all hotels,[14] do not use a CRS. Finally, RMSs that have access to the data outlined in Sec. 4.2 as well as pricing and availability control have a similar level of market penetration. Given the high fragmentation of PMS providers, the low penetration of RMSs, and the CRS market being somewhere in-between, it is hard to imagine how the vast majority of hotel operators could adopt field experimentation, even if a suitable commercial system existed.

The technological challenges do not stop there. The hotel technology stack is quite diverse, and so is the customer journey through different acquisition channels. For example, a user may consult both the hotel website and an OTA, as well as consider product bundles before making the booking decision. The experimentation system has no way of knowing the different customer journeys, and this might violate some of the experiment assumptions. Fortunately, careful experimental design can help mitigate these issues, but some degree of user control is necessary (e.g., to minimize the effect of bookings through opaque channels while an experiment is running). Ultimately, the rise of experimentation systems as products, a certain level of expert user control, and a consolidation of the PMS landscape are likely to all be necessary parts of the solution to the above technological challenges that hinder the adoption of experimentation. But before this happens, it is likely that many large and medium-sized hotel chains will adopt such practices through dedicated products targeting the enterprise customer.

Organizational challenges

A culture of experimentation has been touted as being one in which decision making is democratized. It does not matter what your title is, but rather whether your idea works or does not work. This is of course an oversimplified view of decision making in a complex organization, but it illustrates how foreign the experimentation culture may be perceived by many businesses, including several hotel operators. In addition, in an experiment-first culture, small short-term losses that may arise from running unsuccessful experiments are a necessary step for long-term gains. This trade-off between exploration and exploitation can also feel quite unnatural for conservative managers.

A revenue management team focused on experimentation does not directly make pricing decisions. Instead, they decide which experiments to run by enabling a data science team to conduct these experiments. The results from the experiments are then used to make the final decision, either to implement the tested revenue strategy, or to stop

[13] Exact numbers are hard to estimate, but industry reports can provide an indication: https://blog.capterra.com/opera-vs-maestro-two-popular-hotel-pms-solutions-compared/.

[14] See, for example, the Intercontinental Hotels Group 2019 annual report.

pursuing the strategy, or to follow up by running a new set of refined experiments. This is a significant structural change at the core of the organization, as illustrated in Fig. 4. The first obvious change is the appearance of a data

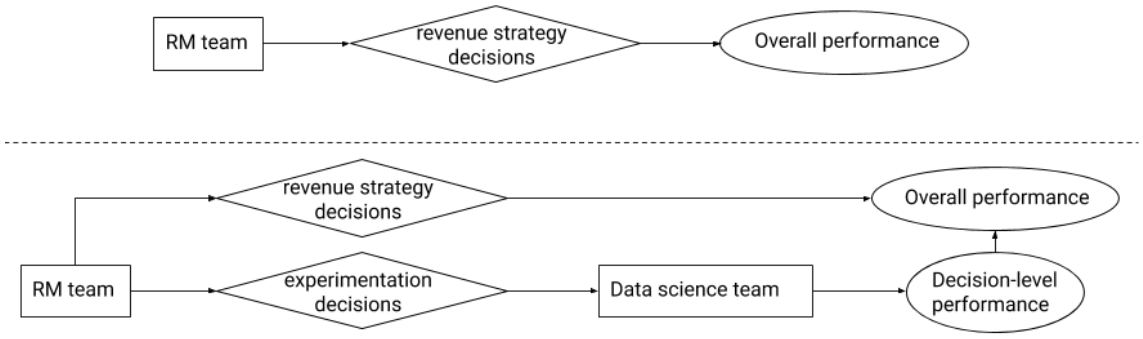

Figure 4: Illustration of how revenue strategy decisions occur in a traditional organization (top) and in an experimentation-first organization (bottom). In a traditional organization, the revenue management team makes revenue strategy decisions based on their experience, external data sources, and results from business intelligence analyses. The quality of these decisions is measured after some (usually long) time period by examining the overall performance of the business. In an experimentation-first organization, the revenue management team makes decisions about which experiments to run. These experiments are setup and analyzed by the data science team, which is then able to assess the quality of each decision separately (after a short time period) and report back to the revenue management team. Realistically, an experimentation-first organization will preserve some of the functions present in a traditional organization, as illustrated in the figure.

science unit. Even if an experimentation platform could enable running efficient experiments, ultimately a data science function needs to work on the experiment parameters and on interpreting the results. This is clearly burdensome for small hotel chains or independent hotels, but data consulting firms could fill this gap for small businesses, in the same way that revenue management is often outsourced. There is another significant change: the revenue management team is no longer directly in charge of some of the revenue strategy decisions and, instead, needs to decide which experiments to run and what to do based on the outcome of these experiments. This is a significant disruption in the role of the revenue management team in the hospitality industry, particularly as larger number of decisions start to rely on experimentation. In large organizations, revenue management teams often include several individuals with backgrounds in operations research and other technical disciplines. But experimental statistics does not figure prominently in most revenue management programs, so that many hotels could find it hard to hire scientists with the required experience to manage a team with an experimentation-first culture.

The above organizational challenges seem hard to overcome, but in reality, they have already been overcome in several large organizations. As illustrated in Fig. 4, the traditional functions will co-exist with the experimentation functions, maybe indefinitely, since some decisions have clear business-level motivations. Therefore, the transition is manageable since the experimentation side of the organization can gain more responsibilities gradually. For small organizations, cultural changes can even be easier, as long as they are catalyzed by the right management and infrastructure that help reduce the cost of adoption.

Another important organizational challenge to experimentation is that the hotel brand or management group might not own certain properties (which are franchised). This bears several consequences. Most importantly, the hotel owners may not agree to participate in an experiment involving many hotels in the brand, or to implement the learnings that result from the experiment. In addition, a brand operating under a franchise model might not benefit from revenue increases directly, and thus has a weaker incentive to optimize pricing strategies for certain properties. Finally, individual hotels may enforce specific business constraints (e.g., maintaining the hotel occupancy at 90% on average), which might make some experiments not viable.

Most of the discussion in this section has focused on the challenges that hotel operators face to embrace field experimentation. Sec. 4.1 pointed to opportunities for RMS providers. In principle, RMS providers are in a good position to extract value from field experiments. Since RMS providers are software or data science companies at their core, they do not face organizational or technological challenges. They, however, face a unique commercial challenge: if clients (hotels) cannot see the value of running field experiments, that will discourage investments in experimentation

infrastructure. It is fair to say that if every client was conducting a field experiment before choosing between different RMSs, RMS providers would have to quickly develop internal experimentation systems to tap into new businesses.

5 Experimental designs

Many of the challenges described in the previous section can be addressed through a careful experimental design. The underlying assumptions necessary for the experiment to be successful need to be verified in each case by the person (or team) conducting the experiment. Each experiment will be different, in light of the question it is trying to address, and depending on the property (or properties) it can use for this purpose. Even when experiments differ in their design and/or implementation, analyzing their results will usually follow well-established practices. In this section, we outline three types of experimental designs that overcome many of the challenges discussed in Sec. 4.3 and are thus well-suited to exploit the opportunities presented in Sec. 4.1. In Sec. 5.4, we highlight well-established empirical techniques that can be used to interpret the results of these experiments and establish causal claims.

5.1 Property splits

Running an experiment on multiple properties, some of which receive the treatment (treated group) and some of which do not (control group), is perhaps the simplest conceptual design. This option is appealing (and feasible) for large hotel brands with many properties in similar markets or for a vendor who is trying to assess the value of a new idea. A simple version of this idea is to use half of the properties for *treatment* (e.g., implementing a new revenue management system) for some time period, while the other half of the properties remain unchanged (called the *control* properties). Precise numbers are specific to each experiment, but as an example, let us consider running such an experiment with 100 properties (50 are treated and 50 are used for control). We assume that the day-to-day standard deviation of `RevPAR` in those properties is 40%, and that there are negligible correlations between revenue fluctuations day-to-day. Assume also that we find a 2% difference in the daily `RevPAR` between treatment and control at the 95% confidence level after 30 days of running the experiment. Most of the examples described in Sec. 4.1 would show differences around this magnitude. When extrapolated to a group of 100 properties, a 2% gain is equivalent to a $2.2M increase in yearly profits for a test that took 30 days to run.

Unfortunately, this simple scenario often does not survive the real-world complexity. The daily `RevPAR` of different properties are hardly ever comparable, and if one is to compare percentage increases instead, it introduces a random variable in the denominator of our performance metric—typically decreasing the power of the experiment. An alternative common way is the use of *stratification*. In the current context, one might group properties according to their `RevPAR` (or alternatively by property type or by amenities) to create homogeneous sets of properties and run several experiments within each set. These individual experiments will have less power than the original experiment, but the combined results will have a higher power than the original experiment. If the experiment treats stay nights which are quite close to the starting date of the experiment, the remaining inventory will vary dramatically across stay nights and across properties (some properties' booking periods are much longer than others). Remaining inventory will often be an important covariate to control via stratification. More complex methods of stratification can be devised, particularly as the number of covariates increases, but discussing those is beyond the scope of this paper.

A naive random split of properties is also subject to suffer from network effects, as discussed in Sec 4.3. Network effects can be minimized by performing a cluster-based randomization, rather than a simple binomial randomization. The difference between the two types of randomization is illustrated in Fig. 5. The *unit of inference* of a cluster-based randomization experiment can either be the cluster itself (e.g., the cluster-level daily `RevPAR`), or the individual elements in the cluster (e.g., the hotel daily `RevPAR`). The former leads to a loss of power, but simplifies the analysis, since the *intra-cluster correlation coefficient* (e.g., the extent to which units in the same cluster affect each other) does not need to be estimated. Because of the added complexity of using clusters and the loss of power, small clusters are usually preferred. Several studies will use an ad-hoc definition of clusters based on intuitive features (e.g., type of hotel, city), and use data-driven estimates of the intra-cluster and Pearson correlation coefficients after the experiment has finished to verify that network effects are indeed small. More sophisticated cluster procedures can be designed based on correlation coefficients of historical data, clustering techniques (e.g., K-means), or propensity score matching [43].

Interestingly, stratification and cluster-based randomization can be combined. One use case of special interest for hotel owners is an experiment in which they want to control for the impact of revenue managers. A hotel chain may have several hotels in a region, which are operated by different revenue managers. Stratification can be used to create individual experiments for each revenue manager, while cluster-based randomization can be used to reduce network effects due to the potential overlapping demand. This is illustrated in Fig. 6. If the actions of the revenue manager in one hotel are influenced by the outcomes in other hotels, there might be network effects that are unaccounted for (which could be mitigated through "revenue manager"-level clusters), but that can be evaluated in a post-hoc analysis and the

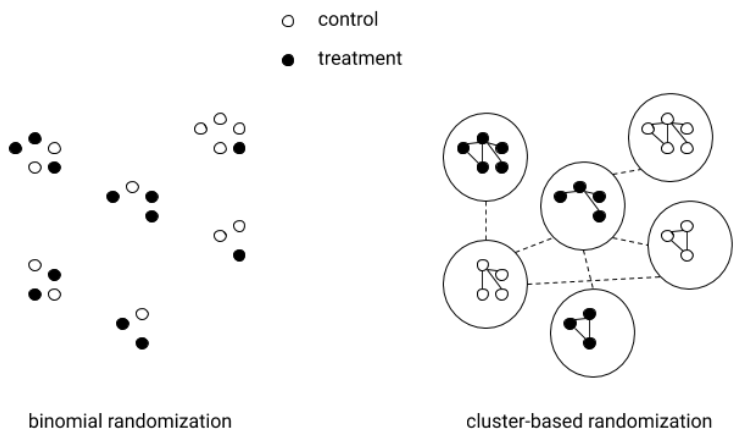

Figure 5: A binomial randomization (left panel) and a cluster-based randomization (right panel). The solid lines represent strong connections (such as shared customers or the same revenue managers), whereas the dotted lines represent weak connections that can carry small network effects.

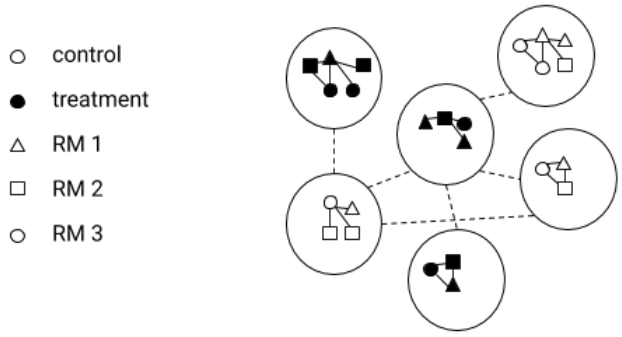

Figure 6: Example of a potential stratification, based on the revenue manager in charge of each hotel, for an experiment that uses cluster-based randomization. In this example, the clusters for each strata are small (size 1 or 2), so that the preferred unit of inference would probably be the cluster level.

affected strata can be eliminated from the final analysis if such correlations are identified. This procedure does not invalidate the experiment as long as it is specified before the experiment starts, but it will potentially require to run a longer experiment.

The process of stratification helps control for relevant *covariates* (i.e., variables that impact the performance measure of interest) and can increase the power of the experiment. However, it does not address biases that may arise from reference effects. As an example, consider the situation with similar properties in our experiment, but the treatment impacts positively the `RevPAR` of half of them, while it impacts negatively (by a similar magnitude) the `RevPAR` of the other half. It is then possible that, by luck, the randomization selects more properties with a positive treatment effect. In this case, we will end up over estimating the treatment effect, even though the effect of the treatment was, in fact, negligible for the entire sample. This problem can be solved by running an experiment in which the treatment and control properties are swapped in the middle of the experiment. For example, during a 30-day experiment, we can use the original assignment for 15 days, while the other 15 days will be under the alternate assignment. One can be more creative with the specific assignment, especially if there are concerns about seasonal effects. In fact, we will discuss this further in Sec. 5.2.

The previous discussion presents several techniques that help address the scientific challenges discussed in Sec. 4.3, while avoiding some of the challenges created by technological limitations. However, it should be clear that there are many things that could go wrong if the experimenter (or the experimentation system) has not taken into consideration all the relevant variables. For this reason, it is always advisable to simultaneously run a closure test (or A/A test) as part of the experiment, to ensure that no important effects that could bias the experiment have been overlooked. For hotel rooms, which are highly differentiable over time, a good way of running a closure test is by dividing the properties into two groups. Stratification, clustering, and any other experimental design features are applied to both groups. In one of the groups, we run the controlled experiment, while in the other group, the randomization is performed as if we were running a controlled experiment, but no hotel is treated. The latter group is used as the A/A test, and no difference should be observed within the confidence intervals (say, with a 95% confidence level). A closure test can also be run before the experiment starts (i.e., during the design phase), but this could miss biases that are time-dependent and only present during the testing period.

This type of field experiment could potentially exploit the opportunities discussed in Sec. 4.1. In addition, this design can be implemented manually most of the time, so that no complex experimentation system is needed to execute the experiment. However, its granularity is typically coarse. Accordingly, it can answer the question "what is the best revenue management system for a hotel group?", but not necessarily "what is the best revenue management system for a specific hotel?" This is not necessarily a weakness, since many questions about revenue strategy need to be asked at the hotel group level. However, it is not suitable to answer questions that are only relevant at the hotel level. For example, when trying to price ancillaries or room attributes, it may not make sense to group several hotels together. For this type of more granular studies, we next present two other types of experimental designs.

5.2 Alternating periods

Not all hotel operators have enough properties to run a meaningful field experiment using the design presented in Sec 5.1. Even for large operators, it might be interesting (and cost effective) to infer the effect of an intervention by using a relatively small set of hotels. The design described in this section expands on the temporal randomization used in the previous section to address reference effects. In this case, the temporal split doubles as a way of reducing reference effects and creating control and treatment samples. Since only a few changes to the revenue strategy are required, the execution of this experimental design can also be done manually for certain applications, even though the execution overhead is significantly heavier than the experimental setup from Sec 5.1 (at least 5–10 times more changes are required in this case). This design can also be used to exploit the opportunities outlined in Sec. 4.1, when only a small number of properties can be used in the experiment. However, as we will see below, certain conditions need to be met, and the experiment duration will generally be significantly longer.

Fig. 7 shows an example of how this temporal split could be implemented. As in Sec 5.1, this temporal split is along

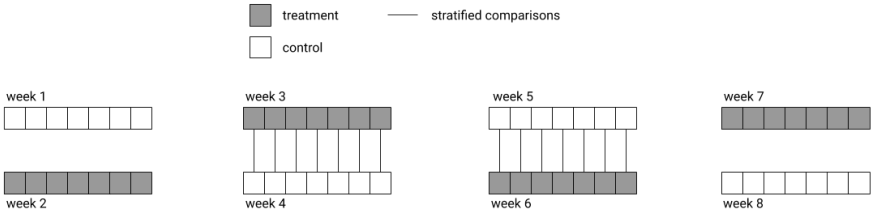

Figure 7: Experiment with temporal split using stay nights. The boxes illustrate individual stay nights. Stratification for each day of the week is included, so that a treated Monday is compared to a control Monday, a treated Tuesday is compared to a control Tuesday, etc. An individual experiment compares `RevPAR` in each night of week 3 with that in each night of week 4. A similar comparison is performed between nights in weeks 5 and 6. Other weeks are used to avoid leakage and as additional controls.

the dimension of a stay night. This means that each stay night will use one revenue strategy for all its bookings. Depending on the length of the booking window, the experiment start date may need to be several weeks before any stay night is treated so that the vast majority of bookings are received when the treatment is applied. However, the techniques introduced in the previous section, as well as the ones that will be discussed in Sec. 5.4, can help handle the heterogeneity arising from different stay nights having different levels of remaining inventories at the beginning of the experiment. This experimental design can be enhanced for operators with two or more properties by running it simultaneously on two properties and alternating the control/treatment assignment between the two properties (i.e., when property 1 is treated, property 2 is used as control and vice versa). This enhanced design provides a strong strategy

for identification of the treatment effect by absorbing both temporal and cross-sectional variations between the treated and control groups.

In the example of Fig. 7, the split unit is at the week level. The split is not random, but rather staggered (1 week control, 2 weeks treatment, 2 weeks control, etc.). This is technically a quasi-experiment, using the terminology from Sec. 2. The setup could instead randomize the treated weeks, but if the experiment is short, the pattern shown in the figure can account for monotonic changes in demand, which for many properties will happen over a one-month period. Stratification can also happen by room type, so that the pattern in Fig. 7 and the alternate pattern can be randomly assigned to different room types, as long as they are different enough so that there is no risk of cannibalization. Stratification by room types may be highly desirable, since different room types in a hotel will often be characterized by different product offerings, and thus yield a different `RevPAR`. Without a stratified analysis, the experiment may be sensitive to the number of reserved rooms of each type (e.g. double vs. superior double) in the treatment and control periods. Effectively, a stratified analysis will rightfully ignore altogether room types with a small number of rooms for sale (because of large statistical fluctuations). If we consider a hotel with 4 different room types with a `RevPAR` standard deviation of 20% day-to-day, then the experiment depicted in Fig. 7, which only uses 14 comparison days, will detect a 5% difference between treatment and control at the 95% confidence level.

The above architecture is careful to stagger treatment and control weeks to avoid reference effects that a monotonically varying demand over a period of a month could cause. If the experiment is run for a longer period, then the treatment and control weeks can be randomized and seasonal effects will disappear on average. Other reference effects might be relevant in this context. For example, if there is a national holiday or a local festival that affects only specific days, this can strongly bias the results of the experiment. Such a concern is often not relevant in large-scale tests discussed in Sec. 5.1, since events will on average affect treatment and control groups in a similar fashion, but it can introduce large biases in this type of experiment. One can remove the days affected by the holiday from the analysis, but any such decision needs to happen at the time of the experimental design with human input. Anomaly detection methods can also be used to automatically flag anomalous days that can bias the experiment, but an understanding of *what* happened is necessary in order to design a suitable experiment.

Another common reference effect that can impact this experimental design is the business mix. Properties often have a variety of rates offered on any given stay night (negotiated, groups, and public). Since negotiated and group rates tend to be lower than public rates, different business mixes can lead to biased results. For some properties, it may be possible (and interesting) to close some of these rates for the experiment period. For others, this may not be possible. Instead, one can use empirical techniques that can alleviate this concern when analyzing the results (e.g., building a *synthetic* control sample or the use of propensity score reweighing); such techniques will be briefly discussed in Sec. 5.4. However, depending on the scope of the experiment, it may or may not be possible to fully avoid reference effects from heterogeneity in business mixes.

The example of Fig. 7 uses treatment/control windows of 2 weeks. One could choose a different time window, but the longer the window, the longer the test will need to run, and it becomes more important to ensure that reference effects are not affecting the results. On the other hand, a shorter window will introduce a risk of contamination from the other periods. It is frequent for travelers to book rooms for more than one night. If a traveler books one night that is treated and one night that is not, then the treatment will affect the demand for the control nights and vice versa. This contamination will be mitigated if we use a longer time window. Ultimately, the experimenter will have to choose an appropriate window length to ensure a low level of contamination. The right value of the time window highly depends on the context and can be data-driven. We note that for certain applications, such as distribution channel optimization, a longer time window is often necessary, since changes in the OTA commission structure cannot be negotiated to follow a staggered pattern with short time periods.

Performing a closure test in this experimental design is extremely important, since the temporal split can create heterogeneous groups, and covariates could have a strong impact on the experiment results. A good way to control for temporal heterogeneity is to build control samples from some room types, (i.e., having a *multi-control* design). An equivalent experimental setup that does not treat any night is then applied on these room types, for the same nights that the experiment runs. This will only be possible if there is no risk of cannibalization, that is, if stimulating or suppressing demand through pricing on one room type does not affect the demand for the other room type. If the hotel operator has two or more properties in different markets, then this type of design can be enhanced by leaving one or more properties outside the experiment. The risk of cannibalization in this case is likely to be lower than the case of a multi-control design with different room types within the same property. Alternatively, one can run a test based on historical data (which was naturally not treated) or, preferably, using periods around the test period (comparing week 1 and week 8 or week 2 and week 7, for example). Another alternative is to use the same weeks during the previous year as an additional benchmark that preserves the same seasonality. Those tests might suggest that biases exist and cannot be

easily understood or controlled for. In that case, this experimental design is not appropriate, and the design described in the next section can be used instead.

5.3 High-frequency price updates

The experimental design described in Sec. 5.2 aims to measure the effect of interventions with high granularity (at the hotel level). Its main challenge, which can be insurmountable for certain properties, is temporal heterogeneity. Hotel rooms are highly differentiable products, and the same room sold for one night can have a very different intrinsic value from the same room sold for a different night. For example, how can we test two revenue strategies on one Winter break period by using only one or two properties? There are not enough properties to apply the design from Sec. 5.1, and not enough days to use the design from Sec. 5.2. OTAs do not support the functionality of splitting customer traffic to show two different prices to different customers for the same night. And even if they did, the interplay with metasearch aggregators may make any such attempt futile. Furthermore, attempting to split traffic through the hotel website (which is possible for certain types of customers) would often break rate parity clauses with OTAs, making the idea of splitting customers infeasible.

A different way of segmenting customers in an unbiased way is to split the inventory according to the two revenue strategies over time. If a hotel is selling 100 rooms for a specific night, 50 of those rooms can be treated, while the other 50 rooms are used for control. To be able to sell both the treatment and control rooms, the time at which each strategy offers its rooms can be randomized. The first strategy might be active for X hours, whereas the second strategy might be deployed for the next X hours. The value of X depends on the context and is preferably low, particularly for properties that sell their inventory fast, often towards the end of the booking period. The exact value of X can be tuned by the experimenter for each property or even for each room type and sets of dates. This is illustrated in Fig. 8. Assuming that

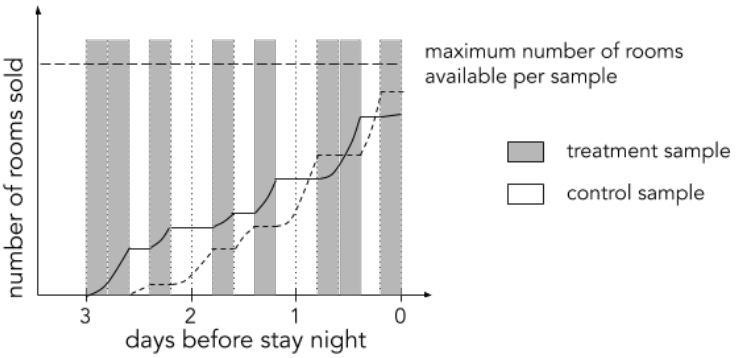

Figure 8: An example illustrating a random split of sale periods between two revenue strategies (treatment and control). In this example, there are 5 periods per day (288 minutes long) and all the bookings are received in the 3 days before the stay night. Illustrative *booking curves* show the number of sales associated with each strategy increasing as the stay night approaches (solid line for treatment and dashed line for control). Selling the inventory from the treatment strategy is not possible when the control strategy is active and vice versa, as evidenced by the flat areas on the booking curves.

customers make their buying decision on the same booking period, this strategy is effective at randomizing the potential customers that would want to buy a room for a specific night. Since the spread in transactional revenue for a single night can be much narrower when compared to multiple nights, this strategy can provide a significant increase in the power of the experiment, which might need to run only for a few stay nights. This strategy may lead to the room being available for sale in some time periods, and not in others, if one of the strategies sells out or blocks some rates, while the other strategy does not. This can be desirable, since availability control is often an important part of a revenue strategy.

Intra-day booking patterns, which often exist, do not affect this experimental setup thanks to the randomization of the periods in which each strategy is active. Bookings from Asia may come earlier in the day, but those potential clients will be exposed to one strategy on some days and to the other strategy on other days. This random assignment will also alleviate systematic biases. The concern of having a different business mix in the treatment and control samples will often not be a problem, since controlling for business mix is an important aspect of a revenue strategy. For example, if half of the inventory is sold at a corporate rate (typically lower than the public rate), it may definitely be part of the revenue strategy. If the strategy that involves the other half of the inventory manages to sell all its inventory at a higher public rate, then this strategy might be smarter about when it should accept corporate bookings.

This experimental design can still be biased by the business mix, but it is often because the formulation of the experimental question is incorrect or the fact that the experimental set up does not properly address the focal question. Consider an experiment that is trying to answer the following question: "Should we price parking at $50 instead of $30?" The answer to this question might be highly correlated to whether the customer is paying a corporate rate. The more corporate bookings, the more interesting it may be to sell parking at $50. But the more corporate bookings, the lower the RevPAR. Ultimately, selling parking at $50 may be interesting during the week days, when the majority of the bookings are corporate, while during the weekend the opposite may be true. This example illustrates that ultimately, even if high-frequency price updates can contribute to a powerful experimental design, careful thought is still needed for each experiment to make sure that the right questions are formulated and that the design is appropriate to answer those questions.

As before, it is wise to build additional control samples, so that a closure test can prove that the experimenter did not overlook key relevant covariates. A natural way to do so for this design is to split the inventory into three groups: one sample for treatment and two samples for control. This can be combined with ideas from Sec. 5.2 in cases where the size of the inventory does not allow creating three samples.

This design is not suited for all purposes. For example, it is hard to see how any question related to distribution channel optimization could be answered with this design. Consider the question of whether paying an extra 2% to an OTA to increase visibility is a good investment. To test this strategy, the OTA would need to allow the property to select when visibility (and the commission value) is increased dynamically, with updates every few hours. OTA contracts currently do not support such a practice. Even if a property is interested in using an experiment to understand whether or not they should be using one OTA, the OTA would likely find out that they are randomly becoming available and unavailable at different times of the day and will stop the relationship for breaching their contract. Similarly, if a property has a fragmented inventory (e.g., all rooms are different), it might be hard to split the inventory into two parts so that both treatment and control samples share similar inventories. This last problem can be overcome by borrowing some ideas from Sec. 5.2.

This design can also be suboptimal for questions which are concerned about an inconsistent customer acquisition journey. Customer acquisition journeys for hotels can be complex. They might start with a mobile Google search, continue a week later by looking at several OTAs and multiple hotels on a desktop computer, and only book two weeks later on the hotel's website (once plane tickets have been reserved). Customers may have become accustomed to seeing price changes through this journey, so seeing different prices may not strongly impact their final decision. However, this highly depends on the hotel's competitive market and on the differences between the two pricing strategies (e.g., customers may feel quite negative about paying 50% higher than the price they originally saw, but they may accept paying a 5% premium). If the experiment is trying to vary the depth of a promotion, then one can easily understand how customers might be disillusioned, and eventually disengaged about first seeing a 20% discount, and later a 10% discount. These are examples of how psychological *anchoring* or *priming* can lead to customer disengagement [44]. The experimenter needs to carefully consider whether an experiment using this design can be biased by such an anchoring effect. If this is the case, one needs to use empirical techniques to flag this effect in the experimental analysis, such as comparing to the previous year's patterns, or looking for anomalous behavior in customers reserving via the hotel website (whose experience can be more easily understood using cookies).

Other technological and commercial challenges may hinder the implementation of this design in practical situations. The execution of this experimental design requires several frequent changes in the revenue strategy, making a manual implementation impractical, and thus an experimentation system is required. Even with an experimentation system, technological challenges may exist in other parts of the hotel technology stack. For example, an experimentation system may change the prices every hour, but the updated price may not be reflected in all OTAs at the same time. The latency of the connections between the PMS (or CRS) and the different channels needs to be verified before implementing this design. Similarly, some OTAs offer products to help customers take advantage of the best price available by rebooking a reservation if the price drops within 7 days from booking. Such products can make this design entirely impractical, except in situations were refundable rates are not available or when the cancellation fee is higher than the maximum price difference tested in the experiment.

5.4 Empirical techniques

Designing a field experiment and extracting information from the data after running the experiment requires dedicated analysis methods. Field experimentation has been developed across several academic disciplines, particularly in econometrics [45], computer science [46], political science [47], and medical sciences [48], so it is hard to find a single book or paper that covers all the relevant techniques. This section does not aim to serve this purpose either. Instead, we introduce some of the empirical techniques that can be most relevant in a hospitality context. We cover techniques that

are useful for pre-experimental analysis, *identification* of treatment effects, and post-experimental verifications that experimental assumptions held during the course of the experiment.

Most generally, the objectives of the analysis of a field experiment will be:

1. determining the main confounding factors,
2. choosing the experimental design to randomize over those confounding factors,
3. for confounding factors that cannot be controlled in the experimental design, establishing the methodology to control for them,
4. verifying that experimental assumptions were not violated during the experiment, and
5. extracting (or identifying) the average treatment effect, or in certain cases, the distribution of the treatment effect over some dimensions of interest.

We have already implicitly identified some of the relevant confounding factors that can impact an experiment throughout this paper. Such factors include: room type, business mix, type of hotel, seasonal or day-of-week effects on demand, network effects, and reference effects. Discussions with the business operations units can be helpful in identifying which of these factors (or others) may be relevant to investigate for a specific experiment. These are often confounding factors because they are typically correlated with `RevPAR`, the most common performance metric to evaluate a revenue strategy. A correlation analysis, either through Pearson's correlation coefficients, Spearman's (more robust to outliers), or any alternative correlation measure, can be helpful in unveiling potential confounders. A cross-correlation analysis between potential confounders can also be used to understand whether some of them are superfluous. For example, it could be that the day-of-week dependence is fully driven by the business mix variation. In that case, considering both day-of-week and business mix in our experimental design will only weaken the power of the experiment. A principal component analysis is a powerful tool that can help identify the most relevant confounders.

The experimental designs presented in Secs. 5.1–5.3 address some of these likely confounders and involve popular techniques to handle confounding factors in the design phase, including randomization, stratified and cluster randomization, and matching [49] (e.g., the idea that some days will only be compared to other similar days as illustrated in Fig. 7). It is the responsibility of the experimenter (with the help of empirical tools) to ensure that these are fit for purpose in each application. Most of these analyses should be made during the experiment design phase and also repeated on the treatment and control samples, to verify that the conclusions from the experiment design still hold during the experiment. For example, one may observe a business mix that is dominated by dynamic rates in the experiment design stage, and decide not to control for business mix during the experiment. In the post-experimental analysis, one may discover that the control sample on average has a similar business mix, but sometimes static rates dominate, as opposed to what has been observed in the design phase. Such an effect would need to be carefully understood and could potentially invalidate the experiment.

Not all confounding factors can be managed through an experimental design. For example, the number of static rate bookings could be hard to predict before the experiment starts, and it is not expected to follow a monotonic behavior. In this case, we may know before the experiment that there will be an imbalance in the business mix between treatment and control. One technique that can help manage this situation post-experiment is by creating a *synthetic* control [50]. In this process, one would weight the control samples in a way that the business mix distribution matches the distribution of the treatment samples as close as possible. For example, if the unit of comparison is the stay night, a weight w_i will be calculated for each stay night so that the business mix distribution in the control sample matches that of the treated sample. Specifically, we can quantify the average treatment effect (denoted t) using the difference in `RevPAR` (denoted r) between treatment and control samples in one of the following two ways:

$$t = \sum_{\text{i}\in\text{treatment days}} \frac{r_i}{N_{\text{treatment days}}} - \sum_{i\in\text{control days}} \frac{r_i}{N_{\text{control days}}} \tag{1}$$

$$\Rightarrow\ t = \sum_{i\in\text{treatment days}} \frac{r_i}{N_{\text{treatment days}}} - \frac{\sum_{i\in\text{control days}} w_i r_i}{\sum_{i\in\text{control days}} w_i}, \tag{2}$$

where Eq. (1) shows a simple difference between the average `RevPAR` across the treatment and control days, whereas Eq. (2) shows the resulting treatment effect after the construction of a synthetic control, where w_is are set to match the control and treatment distributions as close as possible.[15] This procedure is feasible for a small number of confounding factors, but becomes challenging when the number of confounding factors is large.

[15] In the classical synthetic control method, the weights are non-negative and their sum equal to 1. However, several recent extensions have been proposed (e.g., by allowing negative weights).

The verification process of an experiment will at least involve that the assumptions of the experimental setup hold during the experiment period. The design in Fig. 7, for example, is built assuming that the booking window for this property is relatively short, so that all bookings for the comparison weeks are treated in week 2 and beyond. It also assumes that length-of-stay patterns are short enough, so that the contamination between treatment and control samples is negligible. These two assumptions can be explicitly verified before and after the experiment. There will be additional assumptions which are harder to explicitly verify, given their implicit nature. For example, a pre-experimental analysis may not have accurate information about the purpose behind public rate bookings (e.g., leisure vs. business). This might be an important confounder that is not observed by the experimenter. After running the experiment, one can use *placebo* tests to make sure that there is no implicit confounders that affect the results. In the design of Sec. 5.1, one can perform a *cross-sectional placebo test* by randomly splitting the properties into control and treatment (alternatively, one can randomly split only the control properties). One can repeat this process multiple times, creating several unique synthetic datasets, that can be used to verify the absence of biases when the experiment was running, and that the true estimate of the experiment is not an artifact of cross-correlation. A similar process across time, with historical data can be used to perform an *inter-temporal placebo test*.

Typically, the data from a field experiment will be used to first estimate the average treatment effect. In Eq. (1), we showed a simplistic approach to do so (by comparing averages). One can compare the averages by running a t-test (or an ANOVA test for the case with more than two populations). More generally, the process of *estimating* the average treatment effect is often based on a regression specification, as follows

$$y_i = \alpha + \beta \text{Treated}_i + K\text{Controls}_i + \epsilon_i, \tag{3}$$

where y_i is the performance metric of interest (e.g., `RevPAR`) for observation i (e.g., day-property), α is an intercept parameter, Treated_i is a binary variable to indicate whether observation i is in the treated or control condition, Controls_i is a vector of control variables (e.g., confounding factors), and ϵ_i is an i.i.d error term assumed to follow a normal distribution. The average treatment effect is captured by the coefficient β. For robustness purposes, it is typical to estimate Eq. (3) both with and without controls. Using a regression specification allows us to incorporate several modeling assumptions, such as time fixed effects (to capture seasonality), property fixed effects, and explicitly controlling for confounding factors.

A second common identification strategy is based on a *difference-in-differences* specification, as follows

$$y_{it} = \alpha + \beta \text{Treated}_i + \gamma \text{Period}_t + \delta \text{Treated}_i \times \text{Period}_t + K\text{Controls}_i + \epsilon_{it}, \tag{4}$$

where i corresponds to an observation and t to a time period (e.g., before vs. after the experiment), so that y_{it} represents the performance metric of observation i during period t. The binary variable Period_t captures the time period of the observation. The key parameter in Eq. (4) is δ that captures the treatment effect of our experiment relative to the past period. The difference-in-differences benchmarks the performance of the treated samples, pre- and post-treatment, relative to the performance of the control samples, also pre- and post-treatment. The approaches in Eqs. (3) and (4) can be seen as complementary, where the first one relies on cross-sectional variation and the second relies on variation in the time series. When the treatment is alternated and applied at different times (like in the example of Fig. 7), one can enhance the difference-in-differences model from Eq. (4) to a multi-way difference-in-differences estimator [51, 14]. For impactful experiments, it is common to use several empirical techniques, as a way of gaining confidence in the quality of the analysis and in the robustness of the results.

It became common in the field experiment literature to go beyond the average treatment effect, by investigating *heterogenous treatment effects*; either via the distribution of the treatment effect across different dimensions, or by investigating the factors that contribute to the heterogeneity of the treatment effect. Several methods can be used to estimate heterogenous treatment effects. A common alternative is to append an extra term in Eq. (3) (resp. Eq. (4)) which is equal to $\delta_0 \text{Treated}_i \times \mu_i$ (resp. $\delta_0 \text{Treated}_i \times \text{Period}_t \times \mu_i$), where μ_i is the variable that captures the source of heterogeneity (e.g., business mix) and δ_0 is the new estimated coefficient. More sophisticated methods exist, such as causal forest estimates [52].

Estimating heterogenous treatment effects could be of interest to large hotel groups that run experiments across several properties. In this case, one may infer that a specific treatment is most effective for certain set of properties (even if the experiment was not designed to differentiate between property characteristics). Choosing the dimensions for the heterogenous treatment effect depends on the context and often relies on domain expertise.

6 Summary

Over the last two decades, field experimentation has expanded from the academic world into the commercial spotlight by driving a wide range of business decisions. This has led software companies to run concurrently thousands of

experiments with multiple objectives, from informing their product design, to generating stickiness and increasing customer conversion in acquisition processes. Despite its strong focus on data and algorithms, revenue management in the hospitality industry has largely remained detached from this trend: the use of field experimentation is mainly constrained to big hospitality brands, counts a handful of business cases, and has very little public documentation (in academic publications or elsewhere).

The objectives of this paper were to motivate the use of field experimentation in hotel revenue management and to provide a starting point for revenue managers or data scientists who want to use field experimentation to increase revenues. To achieve the first objective, we discussed the economic opportunities of field experimentation in several important revenue decisions faced by hotel managers. We then identified the relevant challenges that the industry needs to overcome for moving towards an experimentation-first culture. While not negligible, these challenges are not insurmountable either, especially in light of the great economic opportunity. Operational burdens of running field experiments would be reduced if a software system (that can serve as an experiment control center) was available to help with the design and execution of experiments. No such tool currently exists, but it has been identified as one of the critical elements that the industry needs to invest on in order to increase the adoption of field experimentation.

To achieve our second objective, three experimental designs, that are particularly well suited to control for common confounding factors, have been introduced. Statistical methods that can be used to evaluate the results from experiments have also been discussed. We note that these designs and empirical techniques are also applicable to other industries in which customers cannot be randomly split in different groups for experimentation, such as brick-and-mortar retail, car rentals, and airlines. We hope that by making these designs explicit and publicly available, the barrier to entry will be reduced for hotel operators interested in field experimentation. Finally, we also hope that this paper will stimulate further publications detailing the setups and results of field experiments in the hospitality industry.

Acknowledgements

The authors would like to thank Kelly McGuire, Pelin Pekgun and Dave Roberts for providing detailed feedback on an earlier version of this paper. We would also like to thank Cindy Heo and Andrew Vakhutinsky for useful comments about the structure and the bibliography. Finally, we are also grateful to Minna Vaisanen for helpful comments and discussions and Nymisha Bandi for providing feedback on the the final draft.

References

[1] M Schrage. *The innovator's hypothesis: How cheap experiments are worth more than good ideas*. MIT Press, 2014.

[2] W-C Chiang, J CH Chen, and X Xu. An overview of research on revenue management: current issues and future research. *International journal of revenue management*, 1(1):97–128, 2007.

[3] R Kohavi and R Longbotham. Online controlled experiments and A/B testing. *Encyclopedia of machine learning and data mining*, 7(8):922–929, 2017.

[4] N Govind. A/B testing and beyond: Improving the Netflix streaming experience with experimentation and data science. *URL: https://netflixtechblog.com/a-b-testing-and-beyond-improving-the-netflix-streaming-experience-with-experimentation-and-data-5b0ae9295bdf*, 2017.

[5] J Overgoor. Experiments at Airbnb. *URL: https://medium. com/airbnb-engineering/experiments-at-airbnb-e2db3abf39e7*, 2014.

[6] P Belobaba. *Air travel demand and airline seat inventory management*. PhD thesis, Massachusetts Institute of Technology, 1987.

[7] S Hormby, J Morrison, P Dave, M Meyers, and T Tenca. Marriott International increases revenue by implementing a group pricing optimizer. *Interfaces*, 40(1):47–57, 2010.

[8] C K Anderson and X Xie. Improving hospitality industry sales: Twenty-five years of revenue management. *Cornell Hospitality Quarterly*, 51(1):53–67, 2010.

[9] B B Oliveira, M A Carravilla, and J F Oliveira. Fleet and revenue management in car rental companies: A literature review and an integrated conceptual framework. *Omega*, 71:11–26, 2017.

[10] P P Belobaba and L R Weatherford. Comparing decision rules that incorporate customer diversion in perishable asset revenue management situations. *Decision Sciences*, 27(2):343–363, 1996.

[11] R Rana and F S Oliveira. Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning. *Omega*, 47:116–126, 2014.

[12] D Koushik, J A Higbie, and C Eister. Retail price optimization at Intercontinental Hotels Group. *Interfaces*, 42(1):45–57, 2012.

[13] P Pekgün, R P Menich, S Acharya, P G Finch, F Deschamps, K Mallery, J V Sistine, K Christianson, and J Fuller. Carlson Rezidor hotel group maximizes revenue through improved demand management and price optimization. *Interfaces*, 43(1):21–36, 2013.

[14] M Cohen, A Jacquillat, and J Serpa. A field experiment on airline lead-in fares. Technical report, Working Paper, 2019.

[15] A Fabijan, P Dmitriev, H H Olsson, and J Bosch. Online controlled experimentation at scale: an empirical survey on the current state of A/B testing. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 68–72. IEEE, 2018.

[16] R Johari, P Koomen, L Pekelis, and D Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

[17] A Deb, S Bhattacharya, J Gu, T Zhou, E Feng, and M Liu. Under the hood of Uber's experimentation platform. *Uber engineering.*, 2018.

[18] G Krishnan. Selecting the best artwork for videos through A/B testing. *Netflix Tech Blog*, 2016.

[19] M Cohen, M-D Fiszer, and B J Kim. Frustration-based promotions: Field experiments in ride-sharing. Available at SSRN 3129717, 2018.

[20] B Halperin, B Ho, J A List, and I Muir. Toward an understanding of the economics of apologies: evidence from a large-scale natural field experiment. Technical report, National Bureau of Economic Research, 2019.

[21] M Cohen, C Fernández, and A Ghose. Empirical analysis of referrals in ride-sharing. Available at SSRN 3345669, 2019.

[22] M Cohen, M-D Fiszer, A Ratzon, and R Sasson. Incentivizing commuters to carpool: A large field experiment with waze. Available at SSRN 3458330, 2019.

[23] J Singh, N Teng, and S Netessine. Philanthropic campaigns and customer behavior: Field experiments on an online taxi booking platform. *Management Science*, 65(2):913–932, 2019.

[24] K J Ferreira, B H A Lee, and D Simchi-Levi. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 18(1):69–88, 2016.

[25] D J Zhang, H Dai, L Dong, F Qi, N Zhang, X Liu, and Z Liu. How does dynamic pricing affect customer behavior on retailing platforms? Evidence from a large randomized experiment on Alibaba. Technical report, Working paper, SSRN, 2017.

[26] M Fisher, S Gallino, and J Li. Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. *Management science*, 64(6):2496–2514, 2018.

[27] V Gaur and M L Fisher. In-store experiments to determine the impact of price on sales. *Production and Operations Management*, 14(4):377–387, 2005.

[28] F Caro, J Gallien, M Díaz, J García, J M Corredoira, M Montes, J A Ramos, and J Correa. Zara uses operations research to reengineer its global distribution process. *Interfaces*, 40(1):71–84, 2010.

[29] C McFarland, M Pow, and J Glick. Quasi experimentation at Netflix. *URL: https://netflixtechblog.com/quasi-experimentation-at-netflix-566b57d2e362*, 2018.

[30] T G Sharma, R Jain, S Kapoor, V Gaur, and A Roy. Oyo rooms: Providing affordable hotel stays. *Emerald Emerging Markets Case Studies*, 2017.

[31] T Cui. Growing our host community with online marketing. *URL https://medium.com/airbnb-engineering/growing-our-host-community-with-online-marketing-9b2302299324*, 2018.

[32] S Srinivasan. Learning market dynamics for optimal pricing. *URL https://medium. com/airbnb-engineering/learning-market-dynamics-for-optimal-pricing-97cffbcc53e3*, 2018.

[33] P Ye, J Qian, J Chen, C-H Wu, Y Zhou, S De Mars, F Yang, and L Zhang. Customized regression model for Airbnb dynamic pricing. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 932–940, 2018.

[34] R L Kaufman, J Pitchforth, and L Vermeer. Democratizing online controlled experiments at Booking.com. *arXiv:1710.08217*, 2017.

[35] B T Öztan, Z van Havre, C Gomes, and L Vermeer. Mediation analysis in online experiments at Booking. com: Disentangling direct and indirect effects. *arXiv:1810.12718*, 2018.

[36] F Parfenov. How we optimized hero images on Hotels.com using multi-armed bandit algorithms. *URL https://medium.com/expedia-group-tech/how-we-optimized-hero-images-on-hotels-com-using-multi-armed-bandit-algorithms-4503c2c32eae*, 2019.

[37] B M Noone. Pricing for hotel revenue management: Evolution in an era of price transparency. *Journal of Revenue and Pricing Management*, 15(3-4):264–269, 2016.

[38] K T Talluri and G J Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Science & Business Media, 2006.

[39] K A McGuire. *The analytic hospitality executive: implementing data analytics in hotels and casinos*. John Wiley & Sons, 2016.

[40] M Hunold, R Kesler, U Laitenberger, and F Schlütter. Evaluation of best price clauses in online hotel bookings. *International Journal of Industrial Organization*, 61:542–571, 2018.

[41] R G Cross, J A Higbie, and D Q Cross. Revenue management's renaissance: A rebirth of the art and science of profitable revenue generation. *Cornell Hospitality Quarterly*, 50(1):56–81, 2009.

[42] L Masiero, C Y Heo, and B Pan. Determining guests' willingness to pay for hotel room attributes with a discrete choice model. *International Journal of Hospitality Management*, 49:117–124, 2015.

[43] P R Rosenbaum and D B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[44] A M Bandalouski, M Y Kovalyov, E Pesch, and S A Tarim. An overview of revenue management and dynamic pricing models in hotel business. *RAIRO-Operations Research*, 52(1):119–141, 2018.

[45] J D Angrist and J-S Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[46] C Wohlin, P Runeson, M Höst, M C Ohlsson, B Regnell, and A Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[47] R B Morton and K C Williams. *Experimental political science and the study of causality: From nature to the lab*. Cambridge University Press, 2010.

[48] J N S Matthews. *Introduction to randomized controlled clinical trials*. CRC Press, 2006.

[49] E A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[50] A Abadie, A Diamond, and J Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

[51] A Goodman-Bacon. Public insurance and mortality: evidence from Medicaid implementation. *Journal of Political Economy*, 126(1):216–262, 2018.

[52] S Wager and S Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.