Proving the performance of a new revenue management system

Kalyan Talluri * Fernando Castejon Begoña Codina Juan Magaz†

November 16, 2009

Abstract

Revenue management (RM) is a complicated business process that can best be described as control of sales (using prices, restrictions, or capacity), usually using software as a tool to aid decisions. RM software can play a mere informative role, supplying analysts with formatted and summarized data who use it to make control decisions (setting a price or allocating capacity for a price point), or, play a deeper role, automating the decisions process completely, at the other extreme. The RM models and algorithms in the academic literature by and large concentrate on the latter, completely automated, level of functionality.

A firm considering using a new RM model or RM system needs to evaluate its performance. Academic papers justify the performance of their models using simulations, where customer booking requests are simulated according to some process and model, and the revenue performance of the algorithm compared to an alternate set of algorithms. Such simulations, while an accepted part of the academic literature, and indeed providing research insight, often lack credibility with management. Even methodologically, they are usually flawed, as the simulations only test "within-model" performance, and say nothing as to the appropriateness of the model in the first place. Even simulations that test against alternate models or competition are limited by their inherent necessity on fixing some model as the universe for their testing. These problems are exacerbated with RM models that attempt to model customer purchase behavior or competition, as the right models for competitive actions or customer purchases remain somewhat of a mystery, or at least with no consensus on their validity.

How then to validate a model? Putting it another way, we want to show that a particular model or algorithm is the *cause* of a certain improvement to the RM process compared to the existing process. We take care to emphasize that we want to prove the said model as the cause of performance, and to compare against a (incumbent) process rather than against an alternate model.

In this paper we describe a "live" testing experiment that we conducted at Iberia Airlines on a set of flights. A set of competing algorithms control a set of flights during adjacent weeks, and their behavior and results are observed over a relatively long period of time (9 months). In parallel, a group of control flights were managed using the traditional mix of manual and algorithmic control (incumbent system). Such "sandbox" testing, while common at many large internet search and e-commerce companies is relatively rare in the revenue management area. Sandbox testing has an undisputable model of customer behavior but the experimental design and analysis of results is less clear. In this paper we describe the philosophy behind the experiment, the organizational challenges, the design and setup of the experiment, and outline the analysis of the results. This paper is a complement to a (more technical) related paper that describes the econometrics and statistical analysis of the results.

*Kalyan Talluri, ICREA and UPF, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain, email: kalyan.talluri@upf.edu

†Iberia Airlines, Madrid, Spain

Key words. Revenue management, airlines, sandbox testing,econometric analysis.

1 Introduction

Revenue management (RM) is a complicated business practice. It involves understanding industry practices, idiosyncracies of each market, customer behavior and preferences, competition, pricing and somehow translating all that experience, wisdom, and information into an operational practice that consistently fetches improved revenues for the firm. In practice it is a combination of analyst skills, managerial knowledge and clever use of data and software tools.

The RM models and algorithms in the academic literature focus on just a narrow aspect of revenue management—the development of models that automate the analysis of the data and convert it into inputs for models which subsequently give optimized decisions (prices or booking controls). In practice firms rarely rely entirely on models or algorithms. A working RM system is usually an undescribable mix of analyst decisions and automation. It is not at all clear then how a RM algorithm, even if it is based on a complicated model, and sophisticated algorithms, would compare to an analyst performing the same task manually based on his or her experience and knowledge of the market.

A firm considering using a new RM model or RM system however needs to evaluate its performance. Academic papers justify the performance of their models using simulations, where customer booking requests are simulated according to some process and model, and the revenue performance of the algorithm compared to an alternate set of algorithms. We argue below that a simulation comparison against another alternate model of the universe is of little use for a firm. The sole concern of the firm is to improve *its* revenues. For this reason such simulations, while an accepted part of the academic literature, and indeed providing some research insight, often lack credibility with management.

Even methodologically, simulations (in the RM context, and the ones in the RM literature) are usually flawed, as they only test "within-model" performance, and say nothing as to the appropriateness of the model in the first place. Even simulations that test against alternate models or models of competition are limited by their inherent necessity on fixing the alternate model for their testing. These problems are exacerbated with RM models that attempt to model customer purchase behavior or competition, as the right models for competitive actions or customer purchases remain somewhat of a mystery, or at least with no consensus on their validity.

How then to validate a model? We want to show causality, that is impute performance to the new method convincingly, which is often a much more tricky exercise than showing correlation, with controversy surrounding even the meaning of "causality" (see Granger [6]). We follow the definition of Hart and Honoré [7]: 'the cause is a difference to the normal course which accounts for the difference in the outcome'.

Showing cause is perhaps the least controversial in an experimental setting, which is the main reason we settled on the "live" testing experiment that we describe in this paper. A set of competing algorithms control a set of flights during adjacent weeks, and their behavior and results are observed over a relatively long period of time (9 months). In parallel, a group of control flights were managed using the traditional mix of manual and algorithmic control (incumbent system). Such "sandbox" testing, while common at many large internet search and e-commerce companies is relatively rare in the revenue management area. Sandbox testing has an undisputable model of customer behavior but the experimental design and analysis of results is less clear. In this paper we describe the philosophy

behind the experiment, the organizational challenges, the design and setup of the experiment, and outline the analysis of the results. Our focus is not so much on the results, the revenue performance of the algorithms, but on how we went about justifying and analyzing the results, that we hope will serve as guidelines for such future live testing. This paper is a complement to (a more) technical related paper that describes the econometrics and statistical analysis of the results.

2 RM Models and Simulations

Most revenue management implementations are based on what is called the "independent class" assumption where the demand for the different fare classes is independent of each other, and more importantly, independent of the controls; the sales are assumed lost when a fare class is closed. This is clearly a rather weak modeling of how customers purchase products, and a number of authors have proposed revenue management using more realistic customer behavior models ([12].

As such models proliferate it is increasingly becoming an issue identifying which is the "right" model. For us a "right" model is one that obtains the firm the most revenue when implemented. This definition avoids judging a model by its elaborateness or complexity (for instance correlations, or allowing dependencies etc.) or generalities, or, one's projection of own behavior or anecdotal evidence. Operational constraints also play a role: to be implementable, a model has to be tractable using current computing capabilities. So an apparently weak model may have as much chance as a sophisticated model if it is more robust, can be run more frequently, or simply is better where it matters.

Modeling customer behavior explicitly is a step in the right direction, but very few RM models in operation (as opposed to proposed in the literature) incorporate competitive behavior. In practice this is a "first-order" factor. Often an analyst will set a price control solely as a response to a competitor's action. Given the difficulties of analyzing models with both consumer behavior and competitor (three players) it is quite likely that RM algorithms will be ignoring some important elements that make up real-world RM practice.

2.0.1 RM System objectives

In our view a RM system, apart from its broad objective of maximizing revenue, ought to have the following characteristics

1. Controllable: analysts will know things no system will know, and hence should be able to allow analyst input.
2. Robust: does not perform badly in any market or circumstance.
3. Adaptable: as we can never forecast certain events, or know certain data (competition controls) the system should therefore be able to adapt (or react) well and quickly.

While RM models have a clear objective function—maximize expected revenue under a stochastic model of demand—what they fail to capture are these "other" aspects of a good RM system. Now, a RM implementation that has been in operation for a while, has some aspects of all the above control, robustness and adaptability, which makes makes evaluating whether a new algorithm is good or bad very difficult. They are also precisely some of the reasons why Monte-Carlo simulations

do not inspire confidence, as these factors are very difficult to simulate. We elaborate further below some further difficulties with RM simulations.

2.0.2 Simulations

Testing RM models then poses a problem. Traditionally, at least in the academic literature, revenue management models and algorithms were tested using simulations. The models generating customer requests in such simulations were more often than not, the exact same model on which the system was based upon. Such simulations are useful for testing performance of algorithms within the confines of a model, but do not give much information about the validity of the model itself.

Even testing based on historic data is not a reliable predictor of generalization as historic data covers a small slice of the relevant market data. One could conceivably run simulations modeling various scenarios and models of customer and competitive behavior (and indeed we mention a few such industry simulations later), but it still does not solve the problem of subjectivity and arbitrariness in choosing one of the models used behind the simulations as the "right" one.

We summarize below some of the criticisms one can level against simulations:

1. Simulations are model based, and the only universally accepted model of customer behavior, as utility maximizers, is too vague to be simulated.
2. One rarely knows what model competition uses and its objectives, so it is difficult to model competitive reactions in a simulation.
3. Unexpected events, by definition, do not follow any model and they play a big role in defining the performance of a RM system
4. Simulations give an unfair, and meaningless, advantage to models that coincide with the one behind the simulations.

We have to admit on the other hand that simulations, while suffering from the aforementioned flaws of arbitrariness in modeling and assumptions, have the advantage of total control over the environment making their design and analysis relatively easy.

3 Context and motivation

The motivation for this study comes from a RM method developed by the authors that explicitly models customer purchase behavior (with a mixture of customer types) and potential market on a flight-date level, and also incorporates other information such as product characteristics and competing offerings. A prototype was developed and refined over a period of two years (from now one we shall refer to this method and its algorithms as *prototype*)[1]. It was at that point that we faced the problem of convincing management on the potential benefits to the firm from using this new method on an operational basis.

Building an elaborate simulation model was not ruled out, but it was not clear (i) how to validate such a simulation model (ii) if results from such as study would ever convince management. So

[1]As the point of the paper is to describe the validation of a RM method rather than the method itself, we do not go into the details of the method.

instead, we decided to take a calculated gamble and test it on live flights. We describe the concept and the setup next.

3.1 Sandbox Testing

A set of competing algorithms control a set of flights, and their behavior and results are observed over a relatively long period of time (9 months). In parallel, a group of control flights were run using the traditional man-machine mix of manual and algorithmic control.

3.1.1 Choice of flights and markets

The objective is to choose a small set of markets. We started off with one market, but eventually the test included ten different markets (city pairs). All the flights were of the point-to-point type. The markets were chosen to be of different types (monopoly, traditional, low-cost competition). The choice of the markets itself was made by the users (the analyst group). We believe it is close to random with perhaps a bias towards the poorly performing markets, as no one is inclined to mess with a top-performing market. Within each market the same user group chose a set of flights; a few complete markets, that is where all the flights in the market were controlled by the test method, were chosen (22 flights overall in one test)

3.1.2 Process

To isolate just model performance and not to get into the cost of denied boardings etc, we decided to use the same overbooking limits for all flights in the test set. The limits were given by the incumbent model. The process of switching back and forth between the incumbent process and the prototype was in fact extremely easy. Figure 1 shows a schematic of the process. The incumbent system outputs the overbooking limits and the curtain settings for the flights. Under normal operations, the incumbent system writes to a directory and another process reads the output and uploads the booking controls to the reservation system, and analysts view and control the settings from then on. The new control system (prototype) is introduced into the process simply by asking the incumbent system to write to a different directory and having the prototype write into the uploads directory in the same format (and with the same overbooking limits) as the incumbent system. So the prototype (and the other systems we eventually ended up testing this way) has the same information as the incumbent system and both systems use the same overbooking limits (set by the incumbent system).

3.1.3 Alternating control and organizational issues

Once the process was put in place, the control for the test flights alternated between the incumbent system and the prototype on alternate weeks. So the incumbent process (the man-machine mix) controlled the test flights for one week (Monday to Sunday) and the prototype took over for the next week. The prototype ran unsupervised after an initial break-in period, so it was a truly automated solution vs. the incumbent process. The idea behind alternating control is that if we do this for a sufficiently long period the differences between weeks would be averaged out. This turned out to be more or less true, with some exceptions as we discuss in a later section.

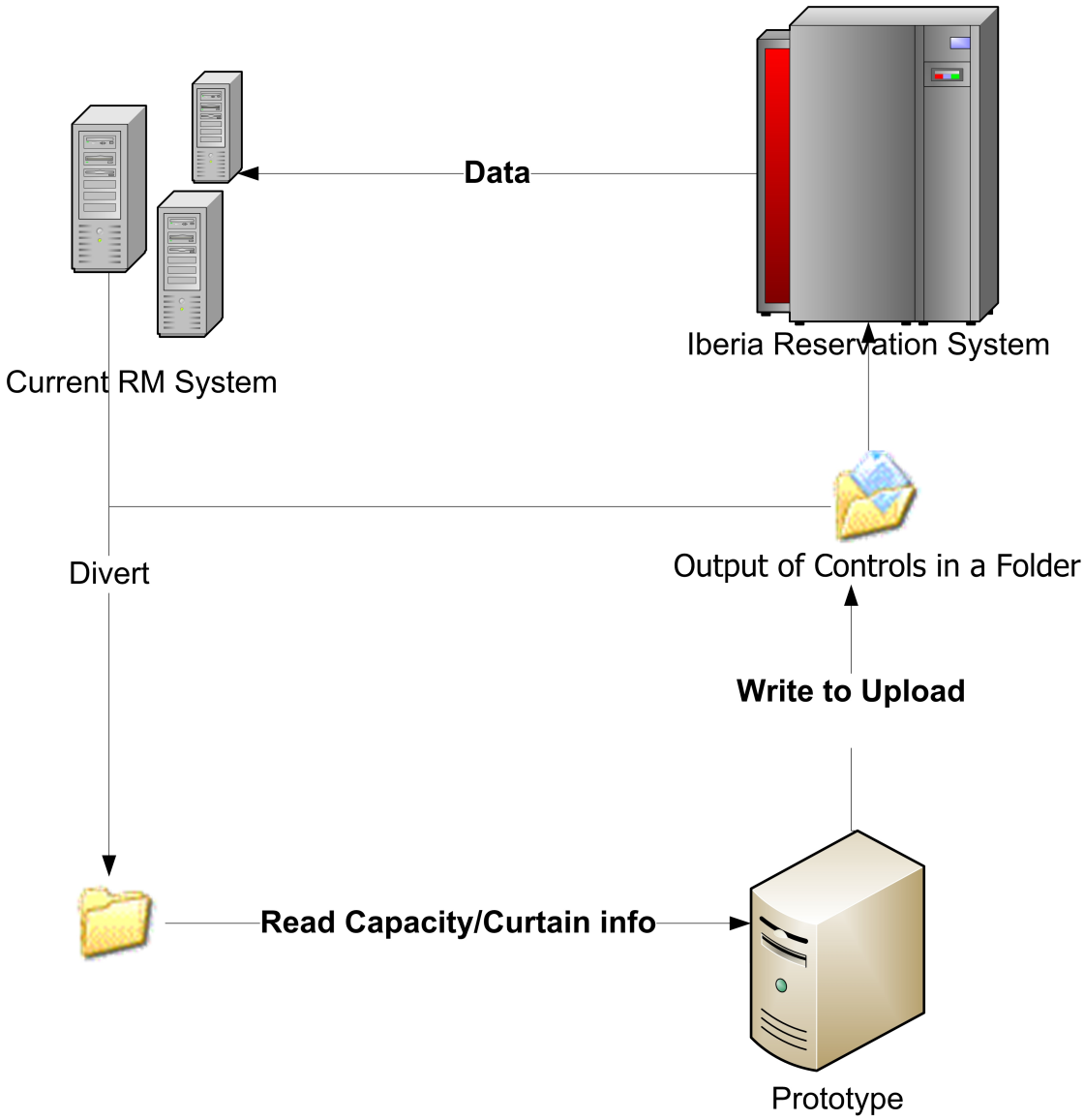

Figure 1: Schematic showing the implementation of the prototype (and other methods).

Organizationally, the biggest challenge was to prevent analysts interfering with the flights during the period controlled by the prototype. As—at least at that stage—analysts were assigned by markets this involved only a few analysts. From a human motivation point of view, just the fact that their work was being tested would have clearly influenced the performance of the analysts. So we do not claim that the results are without the so-called "observer" bias—only that the bias would be on the positive side; as the analysts hopefully were more alert and keen to the observation, their performance would have improved, and hence we compare the prototype with the best of their performance.

3.1.4 Round-trips

One relevant question in the design of the experiment was the issue of round-trips. There are going to be customers who would purchase a ticket starting during one week (controlled by one system) and returning another week (controlled by another system). So the sale is affected by both the systems. Our choice of using Monday-to-Sunday was somewhat influenced by this consideration. While one can never eliminate these reservations, while still keeping a small inter-temporal distance between the systems, we felt that switching over to a new control on Mondays would perhaps minimize the effects the most. Most business customers return to their origin during the week. Moreover most of the restricted fares have a Saturday night stay control, so the leisure customers changing their return date would likely return to origin on the Sunday immediately following the restriction. Groups, such as cruise-line passengers, who are most likely to depart and arrive mid-week were few in the markets that were chosen. Events such as Easter were removed from the test, so holiday-travel that crosses weeks was also minimal in the final results.

3.1.5 Advantages and complications

Compared to simulations, sandbox testing has some important advantages,

1. Tests what we are looking for (revenue performance) over an extended period of time.
2. Not based on *any* model or assumptions.
3. Based on actual flights and markets.
4. The testing is holistic in the sense that it puts an automated model through all the real-world factors and complications, and compares system (analysts, information) as a whole.
5. It tests the model's reaction capabilities, which is an oft-overlooked aspect of RM systems.
6. Gives a deep and fully credible understanding of what works and doesn't work in the design of RM models.
7. Allows the R&D department to continuously improve the models and algorithms.
8. The results are tangible and believable.

Sandbox testing thus has an undisputable model of customer behavior but the experimental design and analysis of results is less clear. Indeed what we have encountered is that all the work that goes into building a Monte-Carlo simulation has now moved to giving careful thought about the experiment, and in an exhaustive analysis of the results.

4 Literature Review

For the background on revenue management and dynamic pricing, we refer the readers to the books Talluri and van Ryzin [13] and Phillips [10]. Talluri and van Ryzin [12] studied RM explicitly modeling customer choice behavior by a tractable discrete choice model and this has led to much recent research along these lines.

Almost every paper on revenue management presenting a new method or a variation on an existing model or algorithm has some Monte-carlo simulations justifying the performance of the method. Most of these simulations work within the model setting (that is, do not test out-of-model performance). We note here however a series of simulation experiments conducted at MIT called PODS by Belobaba et al. [1], [2]. These simulations are distinguished by the fact that they have an explicit model of customer behavior, and moreover, in the later part of the studies, include competition. So in effect, they are trying to study various effects of customer behavior and competition, but the main focus there is to study the performance of the algorithms.

We are not aware of published literature that describes explicitly the kind of sandbox testing that we describe here, but from our industry conversations, we believe that a number of airlines would have conducted similar studies internally without publishing the results or experiments. We believe our contribution is formalizing the experimental design and the analysis of the results.

The published literature on experimental design and statistical analysis of results in other contexts (especially drug testing) is of course too vast to fully describe here. Cox [3] and Cox and Reid [4] provide much of the background statistical context for experimental design. In the language of experimental design ours is a prospective longitudinal, or a cohort, study.

Our objective in this study is not just to conduct an experiment but also find a causal relationship between the experiment's results and the treatment. This is more tricky. An every statistician knows, correlation is distinct from causality, and determining the latter is an order of magnitude more difficult than the former. This difficulty is captured by the *fundamental problem of causal inference*: it is impossible to observe the effect of more than one treatment on the same unit at the same time. In our context, this simply means we cannot control a flight on a particular day by method A and method B at the same time. This is at the heart of our choice of a longitudinal alternating-weeks experiment.

The literature on causality and its testing is as vast as that on experimental design. We refer the reader to a recent survey of Pearl [9] for background material. Our approach can be said to be close that of the Rubin's causal model of potential outcomes (Rubin [11]; see also [8]) using structural equations as outlined in Pearl [9].[2]

5 Results for four systems on alternating weeks

As we mentioned earlier, the burden of analysis in sandbox testing shifts to the design of the experiment and analysis of results. Since we can control a resource by only one method at any given instance of time, both are serious issues. The results we describe below are actually for a test comparing four different RM methods, the incumbent system, prototype, and two different systems from two vendors. The general idea is to compare revenue performance (revenue per available seat)

[2]There is an alternate definition of causality based on predictability, called Granger causality [5], that has found acceptance in econometrics but we discard as inapplicable in our context.

for this year compared to the same week the year prior.

5.0.6 Difficulties

The main problem with analyzing data is that there could be large differences across weeks, for the same flight. Even though the initial idea was that such unidentified variations would average out by running it for a sufficiently long period of time for alternate weeks, it so turned out that even 9 months of testing, divided over four controls gives only 9 weeks of data for each, and this is not sufficiently large to average out inter-week differences. Alternate weeks are reasonably close to be comparable, but two flight-days four weeks apart are just not comparable as there is a fundamental difference in the underlying demand.

This forces one to bring in the comparable week the year prior to minimize the differences between weeks. As one can see from the data, this greatly reduces out the variance (thus explaining the difference), but there are problems still. Even taking ratios of revenue over comparable weeks (shifted 52 weeks prior), for apparently no identifiable reason—no change in the control, competitive landscape, or flight timings, or events falling differently—there is significant variance. One could only attribute these changes to two things: a different macroeconomic landscape, or a different price structure, especially as prices change frequently in the airline industry.

Of course, there could be identifiable reasons also. For the flights in question the competitor flights were exactly the same, and there were only minor shifts in the flight timings. However, some major holidays, such as Easter weekend (which was the only such for this test period), fall on different days each year, and have to be accounted for. Our solution is to delete such weekends entirely from the test and compare performance during that week in isolation, separately.

Year-over-year comparison is standard in the industry (in all businesses in fact), and hence likely to be accepted by management, so this is one reason we made this choice. However, in a rapidly changing industry with frequently changing price structures, such comparison is fraught with difficulties, and one has to make a careful analysis of the results.

It so turns out that testing over *all* the flights in a market is not valuable as it is very difficult to extract the natural inter-week variation compared to the prior year. The markets where there was at least one control flight which had the same incumbent control throughout however captures the inter-week variations. We compare the revenue performance of the test flight vs. a control flight which was experiencing the same demand variations (compared to the year prior for the same day) and hence comparable. (This however raises some questions of cannibalization that we address shortly.)

Although the results are not the point of the paper, we state them below to show how the perspective can change with a raw comparison vs. a comparison with a control flight during the same week.

5.0.7 Results

In Table 1 we give the revenue performance for the total periods controlled by each one of the methods. It appears there are big differences in revenue performance, and if one looks carefully, load factors as well as average fares. All the methods have gained, including the incumbent method. This could be attributed to a general improvement in the markets or changes in the price structure.

Current year:	Incumbent	System 0	System 1	System 2
Capacity (seats)	177,145	195,059	156,656	195,059
Demand (pax)	121,991	128,856	105,244	130, 589
LF (%)	68.9	66.1	67.2	66.9
Avg. Fare (€)	88.2	89.9	84.6	88.8
Revenue (1000 €)	10,757	11,587	8,908	11,592
Avg. rev/seat (€)	60.7	59.4	56.9	59.4
Year prior:				
Capacity (seats)	178,900	197,050	158,240	197,050
Demand (pax)	105,685	113,828	89,496	116,024
LF (%)	59.1	57.8	56.6	58.9
Avg. Fare (€)	88.3	89.1	85.0	91.1
Revenue (1000 €)	9,334	10,147	7,610	10,568
Avg. rev/seat (€)	52.2	51.5	48.1	53.6
Var. Avg. rev/seat (%)	16.4	15.4	18.3	10.8

Table 1: Revenue results compared to the year prior for incumbent, prototype and two other systems.

However, taking into account the variation of the rest of the flights on the routes (the numbers are only for markets where at least one flight was controlled by the incumbent system throughout) the numbers, in Table 2, paint a different picture. The variation across the different systems has

	Incumbent	System 0	System 1	System 2
Var. Avg. rev/seat (%)	16.4	15.4	18.3	10.8
Var. Avg. rev/seat (%) flights out of the trial	14.2	9.4	16.5	9.6
Differential rev/seat gain	2.2	6.0	1.8	1.2

Table 2: Revenue results compared to the out-of-trial flights in the same markets.

moderated somewhat. The flight that was out of the test during the same week also experienced an improvement in revenue by a significant amount. This flight captures all the unidentified changes in the demand.

Note that only the incumbent system had manual attention. All others were nearly completely automated.

5.0.8 Cannibalization

The natural question that arises, when comparing against a within-week out-of-trial control flight is the role played by cannibalization. It is quite plausible that the test method performed well because it cannibalized demand from the airline's other flight during the day.

Cannibalization can take many forms, especially in the RM context. It can also be confused with genuine market growth or demand stimulation. We therefore identify four forms of cannibalization specific to RM.

1. Demand cannibalization: The test flight prices low and just takes away aggregate demand from the out-of-trial flight. The easiest form to test.
2. Mix cannibalization: Demand apparently is unaffected, load factor proportions are as before, but the test flight alters the mix of the out-of-trial flight (Example: Test flight takes business passengers and gives back leisure passengers to the out-of-trial flight).
3. Consumer surplus cannibalization: The test flight increases its demand by extracting more consumer surplus.
4. Competition cannibalization: The test flight is stealing demand from the competition. In other words, the demand increases comes from increasing market share (say in the time-band of the test flight) rather than from the out-of-trial flight.

The last two forms of cannibalization do not really deserve to be called cannibalization (as they may be seen as the objective of RM), but we include it here to identify the results. One can also dispute the benefits of the last form, competition cannibalization, as it might be detrimental in the long run.

It is quite possible also that the test flights are shifting demand across weeks. However, given the alternatives, and as changing the day of travel is costly to the consumers, we discarded this possibility in our testing.

Finally, one should not forget the possibility that the out-of-trial flight is actually cannibalizing from the test flight, and the performance of the test flight would be higher under "normal" circumstances. This can happen for instance if the incumbent system is pricing the out-of-trial flight too low and the test flight is seeing low demand because of this.

All these above points are to clarify the various possibilities, and to point out the difficulties of testing for cannibalization. Given the nature of the problem and the number of unobservable factors, one can only provide statistical evidence that a certain phenomenon is, or is not, happening. We outline next the nature of such statistical and econometric tests that we performed to analyze the results.

6 Econometric and statistical analysis framework

Now, given the experiment and the results, we want to justify two things: (a) show that the introduction of the new RM method or process caused the differences in the results (b) identify cannibalization, if any. For this reason, we need to specify the universe behind the experiment, some baseline and causal assumptions, and the structural equations we believe are driving the results.

6.1 Universe and causal assumptions

Our universe begins with the notion of potential market demand. This is the total population that has a a reasonable probability of wanting to take a flight in a given market on a particular day. We assume (exogenously) that the market demand is independent across days and across markets. We denote this (unobservable) market demand by M_t for day t. Within this demand there could be a mix of different types (say with different willingness-to-pay for travel).

M_t is shared by the flights in the market. For simplicity (in describing the model) we assume there are two flights of our airline each day (call them flight f and $\bar{f}$ to stand for test flight and

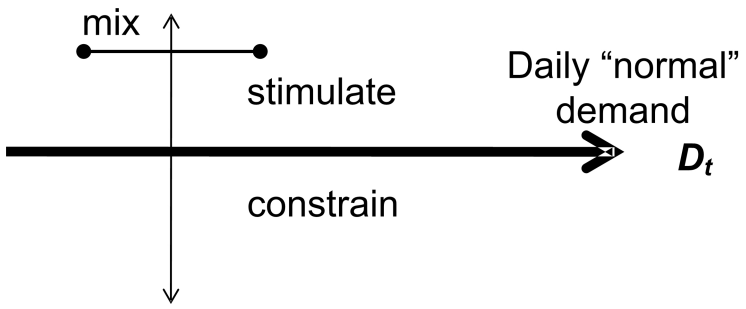

Figure 2: Influence of a new RM system on demand over "normal" demand for a set of daily flights.

out-of-trial flight respectively), and a bunch of competing flights that we group together as a single alternative c. Like wise, let w represent a week when both the flights are controlled by the same incumbent system, and $\bar{w}$ represent a week where one of the flights was controlled by an alternate system (we can make the index by day, but we use weeks as units for simplicity). We refer to them as normal week and test week respectively.

Under "normal" circumstances, say with an incumbent system and process under steady-state, we represent D_t as the daily demand that our airline sees. So this is the fraction of the market demand the airline captures on any given day. Compared to this "normal" demand scenario, a new "treatment", that is a new RM system changes either the mix of the demand, stimulates or restrains the demand.

Our causal assumption is that a RM system or process influences the overall demand that a particular flight sees as well as the mix of the demand that it sees. This assumption is illustrated in Figure 2 assuming a new RM system controls all the flights in the market. So a new RM system can potentially improve (or decrease) revenues, while keeping the same load factors.

Our assumption about system performance (measured say by revenue per available seat) is that it is caused by three factors: (i) underlying market demand, (ii) RM system or analyst skill in extracting consumer surplus (changing the mix), and (iii) the RM process stimulating demand. The first factor is outside the control of the system and should be separated from underlying performance. We want to identify the relative performance of a new system on f in its effective manipulation of the second and third factors. A complication is that we also want to check if a new system is improving to the detriment of the system controlling the out-of-trial flight $\bar{f}$.

We believe we cannot avoid keeping an out-of-trial flight as we need a control for the (unobservable) underlying daily market demand.

6.2 Structural Equations

An econometric model has to be sufficiently detailed to capture the effects of factors we are interested in (such as system performance and cannibalization) and at the same time parsimonious enough to be estimable from observed data. With this in mind, we make our model of demand and RM system performance as simple as possible.

Let I represent observed RM system performance measure (revenue per available seat), and α

true revenue management skill or performance. Then for two flights f and $\bar{f}$, the relationship is given as

$$
\begin{aligned}
M_{\bar{w}} &= \lambda M_{\bar{w}-52}\epsilon_{1\bar{w}} && (1)\\
I_{\bar{w}f} &= \alpha\gamma_f M_{\bar{w}}\epsilon_{2\bar{w}} && (2)\\
I_{\bar{w}\bar{f}} &= \alpha\gamma_{\bar{f}} M_{\bar{w}}\epsilon_{3\bar{w}} && (3)
\end{aligned}
$$

where γ_f represents the normal share and mix of the daily demand. We assume $\gamma_f + \gamma_{\bar{f}} < 1$, with the understanding that competition takes a share of the rest of the market demand. These quantities are of course highly dependent on the RM system controlling both the flights. One could break up γ_f into a demand factor and mix factor, but as we believe we cannot estimate them, at least with this experiment, we combine them into one. The total performance of the system on a particular flight then is $\alpha\gamma_f$.

We are not interested in the values of the parameters per se, but rather, whether, if we introduce a new RM system for flight f, does it change $\gamma_{\bar{f}}$ by a significant amount. So the role of γ_f is to capture flight specific demand and mix.[3]

Equation (1) represents the evolution of the market demand.[4] We assume it is linked to the demand of the same week of the prior year (52 weeks prior, represented by $t-52$). The ϵ's are (multiplicative) random errors.

So now, suppose we introduce, during a certain week w, a new RM system on flight f. The equations for this week would change to:

$$
\begin{aligned}
M_w &= \lambda M_{w-52}\epsilon_{1w} && (4)\\
I_{wf} &= \alpha'\gamma_f^{'} M_w\epsilon_{2w} && (5)\\
I_{w\bar{f}} &= \alpha\gamma_{\bar{f}}^{'} M_w\epsilon_{3w} && (6)
\end{aligned}
$$

So, the new control can potentially affect the γ's (possibly increasing the share compared to the competition).

The cannibalization question now is whether $\gamma_{\bar{f}}^{'}$ is significantly different from $\gamma_{\bar{f}}$. Once we resolve this question in the negative, we can then ask the performance comparison question: whether $\alpha'\gamma_f^{'}$ is significantly different from $\alpha\gamma_f$. If the first question is positive, the results would be inconclusive, at least when $\gamma_{\bar{f}}^{'} < \gamma_{\bar{f}}$.

6.3 Heuristic Justification

In Talluri et al. [14] we estimate the simultaneous dynamic equations (5) and (6) using tools from econometrics. In this paper, we only give a heuristic argument to justify the performance comparison made in Table 2. Our intention here is to show also why the system is estimable in the first place, and also bring out an important assumption we are making on the underlying demand.

[3]For instance, business passengers might prefer a early morning flight and would not shift easily to an afternoon flight, so a morning flight might always have higher RM performance measure even though load factors are the same and both flights are controlled by the same system.

[4]One can make this more complicated, but as this equation captures seasonality effects we believe it is sufficient for our purposes. We remove all data points if the week this year or the prior year has major identifiable events, such as Easter.

We work with logarithms of the equations and compare the equations pairwise ignoring all error terms. Comparing Equation (2) vs. Equation (3)

$$\ln(\frac{I_{\bar{w}f}}{I_{\bar{w}\bar{f}}}) = \ln(\frac{\gamma_f}{\gamma_{\bar{f}}})$$

and comparing Equation (5) with Equation (6)

$$\ln(\frac{I_{wf}}{I_{w\bar{f}}}) = \ln(\frac{\alpha'}{\alpha}) + \ln(\frac{\gamma_f^{'}}{\gamma_{\bar{f}}})$$

and Equation (2) with Equation (5)

$$\ln(\frac{I_{\bar{w}f}}{I_{wf}}) = \ln(\frac{\alpha}{\alpha'}) + \ln(\frac{\gamma_f}{\gamma_f^{'}}) + \ln(\frac{M_{\bar{w}}}{M_w}). \tag{7}$$

To eliminate the last term we compare Equation (1) with Equation (4) for this year and the year prior

$$\ln(\frac{M_{\bar{w}}}{M_w}) = \ln(\frac{M_{\bar{w}-52}}{M_{w-52}}) = \ln(\frac{I_{(\bar{w}-52)f}}{I_{(w-52)f}}) \tag{8}$$

where the last equality is from comparing Equation (2) with Equation (5) for the year prior. This also brings to light an assumption that we are making in Equation (1): we are assuming that the *ratio* of the total market demand over the weeks w and $\bar{w}$ are the same for this year and the year prior. We acknowledge that this is an important assumption, but is nevertheless weaker than saying the demands are the same across years. We need it however so we can eliminate the ratio in each of the terms using the prior year RM performance for the weeks.

Now the cannibalization question can be settled by comparing Equation (3) with Equation (6)

$$\ln(\frac{I_{\bar{w}\bar{f}}}{I_{w\bar{f}}}) = \ln(\frac{\gamma_{\bar{f}}}{\gamma_{\bar{f}}^{'}}) + \ln(\frac{M_{\bar{w}}}{M_w}).$$

Once this is settled we can extract relative performance from Equation (7), comparing $\alpha'\gamma_f^{'}$ with $\alpha\gamma_f$.

While the above reasoning is heuristic, on an average basis, it is what lies behind the comparisons in Tables (1) and (2) (without the cannibalization question). They can be seen as comparing

$$\ln(\frac{I_{wf}}{I_{(w-52)f}}) = \ln(\frac{\alpha'}{\alpha}) + \ln(\frac{\gamma_f^{'}}{\gamma_f}) + \ln(\frac{M_w}{M_{w-52}})$$

and replacing $\ln(\frac{M_w}{M_{w-52}})$ by its proxy $\ln(\frac{I_{w\bar{f}}}{I_{(w-52)\bar{f}}})$ assuming cannibalization has been checked to be negligible.

The equations have to be estimated using daily data of course, and this is what we do (under slightly more complicated dynamics) in Talluri et al. [14].

7 Final comments

In this paper we describe an experiment for testing RM performance on a set of live flights (sandbox testing). With very few assumptions and in actual market conditions it evaluates the performance

of a new system against an incumbent system. We make a case for not using simulations, especially when involving competition and customer behavior and unpredictable events. We provide the background of the testing framework and the design of the experiment, that we hope would be useful for practitioners contemplating a similar question. Finally, we provide a model of RM performance for evaluating the results using dynamic structural equations.

References

[1] P. P. Belobaba. PODS results update: Impacts of forecasting on O-D control methods. In *1998 AGIFORS Reservations and Yield Management Study Group Symposium Proceedings*, Melbourne, Australia, 1998.

[2] P. P. Belobaba. Revenue and competitive impacts of O-D control: Summary of PODS results. In *First Annual INFORMS Revenue Management Section Meeting*, New York, NY, 2001.

[3] D. R. Cox. *Planning of Experiments*. Wiley, New York, NY, 1958.

[4] D. R. Cox and N.Reid. *The theory of the design of experiments*. Chapman and Hall/CRC, Boca Raton, FL, 2000.

[5] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[6] C. W. J. Granger. Testing for causality. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[7] H. L. A. Hart and A. M. Honoré. *Causation in the Law*. Oxford University Press, Oxford, UK, 1959.

[8] P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

[9] J Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[10] R Phillips. *Pricing and Revenue Optimization*. Stanford Business Books, Stanford, CA, 2005.

[11] D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[12] K. T. Talluri and G. J. van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, January 2004.

[13] K. T. Talluri and G. J. van Ryzin. *The Theory and Practice of Revenue Management*. Kluwer, New York, NY, 2004.

[14] K.T. Talluri, F.Castejon, B.Codina, and J. Magaz. A framework for sandbox-testing a new revenue management system with an econometric analysis of the results. In preparation.