II International Conference on Industrial Engineering and Industrial Management
XII Congreso de Ingeniería de Organización
September 3-5, 2008, Burgos, Spain

Extension of Yield Management based Decision Support Systems to continuous and hybrid environments

Alfonso Durán Heras[1], Isabel García Gutiérrez[1], Esmeralda Giraldo Casado[1]

[1] Ingeniería de Organización. Escuela Politécnica Superior. Univ. Carlos III de Madrid. Avda. Universidad, 30. 28911 Leganés, Madrid. alfonso.duran@uc3m.es, isabel.garcia@uc3m.es, esmeralda.giraldo@uc3m.es

Abstract

Yield Management based Decision Support Systems are generally found in discrete, off-line environments, in which the time lapsed between the reservation and the execution date is measured in days or weeks. This paper analyzes their applicability when the process involved is continuous, thus decisions have to be taken in real time, with reservation / allocation taking place only seconds before the execution time, as when allocating customer calls to operators in a call center. Some conceptual differences are analyzed, along with the modeling activities required to render YM-type approaches applicable in those environments. Given the short decision times, the DSS must interact in closed loop with the environment. The forecasting subsystem must thus be fed automatically in real time with the demand evolution; the perishable asset availability schedule and the current reservations backlog must also be updated without human interventions, and all these subsystems must be tightly integrated with the YM based DSS subsystem.

Keywords: Yield management. Continuous environments. DSS. Dynamic systems

1. Utilization of Yield Management algorithms in Decision Support Systems

Yield Management (YM), initially born in the airline industry but currently utilized in diverse service sectors including hotels and car rental, is an approach aimed at maximizing the revenue obtained from a stream of potential customers, with different Willingness To Pay (WTP) or price sensitivity, when resources available to serve these customers are fixed and perishable. It is particularly applicable when variable (volume-related) costs are negligible compared with fixed costs, and thus maximizing revenue results in maximized profits. In airlines, the seats available for a given flight are fixed, the contribution potential of these seats is lost if they are not used (thus they are "perishable"), the costs are nearly independent of the actual number of passengers, and reservations are made by a sequence of heterogeneous customers.

YM splits customers into classes according to one or several criteria (e.g. anticipation with which the reservation is made, reservation channel, customer traits such as age...). These classes (and thus the classification criteria) must allow at least partial price discrimination, i.e., the producer must be able to sell essentially the same service to customers in different classes (thus called "rate classes") at different prices, while preventing customers in one segment from gaining access to the prices offered to the others. Ideally, customers in some "rate classes" should have predictably higher WTP than others, thus guiding the differential pricing; that inter-class WTP variance is not, however, a prerequisite, provided intra-class WTP variance does exist.

Conventional YM algorithms then assist producers in either establishing the prices for each rate class or assigning existing fixed resources (e.g. airline tickets in a given flight or hotel rooms for a given date) to each rate class. For practical applications, these YM algorithms are built into computerized Decision Support Systems (DSSs), the most famous of which was developed by American Airlines, but have since spread into other sectors, such as car rentals (National Car Rental Systems) and hotels (Marriott) (Baker and Collier 2003; Elliott, 2003; Chiang et al, 2007).

These YM-based DSS must include a tri-dimensional demand forecasting module. The three dimensions are: the execution date (e.g., the date at which a hotel room will be used), the time remaining until the execution date, and the demand rate class, subpopulation, or any other proxy for the readiness to pay; each cell contains the forecasted number of customers of that rate class that will make a booking for a given execution date with that anticipation. Based on that forecast, the perishable asset availability schedule and the reservations backlog, the YM module of the DSS can suggest the course of action that is more likely to maximize the yield.

2. Continuous and hybrid environments

Nearly all current YM-based DSSs focus on discrete, off-line environments, in which the time lapsed between the reservation date and the execution date is measured in days or weeks. However, in many settings, the prerequisites for YM applicability (fixed capacity, no stocks, heterogeneous demand, low variable cost) are met, but the process involved is continuous, thus decisions have to be taken in real time, with reservation/allocation taking place only very shortly before the execution time. Some business processes are hybrid: some of the decisions involved are "discrete YM" while as others would fit in the "continuous YM" category.

A representative example of such a hybrid environment is a call centre handling hotel reservations, since room reservation belongs to the conventional, "discrete YM", while as call queues at the call centre can be construed as a "continuous YM" environment. The automated-call-distribution technology utilized (Pietraszek and Ramchandran, 2006) might segment incoming calls according to yield potential, e.g. through calling number identification and automatic lookup in the Customer Relationship Management (CRM) database. In that case, the process of assigning operators to incoming calls can be modeled as a YM problem: the waiting calls of various yield potentials become the rate classes, operators are the perishable assets and the probability of a reservation translates into the probability of answering a waiting call before the caller hangs up. Nair and Bapna (2001), while studying Internet Service Providers (ISP), highlight that they share many traits with traditional YM users. However, this is a continuous environment, in which service request and service delivery are nearly simultaneous, which they model as a continuous time Markov Decision Process.

Due to the business potential of the application of YM based DSS to these continuous or hybrid environments, the "Advanced DSS for hotel management" multi-year, multi-centre research project approved in 2005 by the Spanish Ministry of Education[*] contemplates "... the

[*] This work stems from the participation of the authors in a research project funded by the Spanish Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2004-2007 (MEC-DGI-SGPI), reference DPI2005-09132-C04-04, title Sistema avanzado de ayuda a la toma de decisiones para la gestión hotelera.

extension of the yield management approach to continuous environments with short decision time spans. That extended or generalized approach will be applied, within the reservation process, to other resources besides the rooms, such as the operators in the call centers ...".

3. Continuous environments typology

An initial analysis and modeling exercise for continuous environments from the perspective of exploring the applicability of YM algorithms led to the identification of two basic categories of continuous environments: with or without interactive negotiation.

This key distinction stems from the "autonomy" in the behavior of the "customers". In traditional YM environments, the producer sets the price for each rate class, and then customers for each rate class exercise their autonomy in either buying at that price or not (or try to circumvent the producer's segmentation to access a better price). That cornerstone implicit "customer autonomy" assumption forms the basis for existing YM algorithms, and for the corresponding demand forecasting requirements.

However, in the timeframes found in continuous environments human customers do not generally take price-based acceptance/rejection decisions ("hanging up the phone", which is a different kind of decision, is discussed later). In YM situations, in which a number of perishable resource units must be allocated to specific customers, the multi-party dialogue required can not generally be handled in seconds with human customers. This might, however, change through the eventual generalization of the use of software agents as "purchasers", authorized to commit payments (in whichever type of "currency", not necessarily financial) on behalf of their "principals". Some promising applications for that approach include continuous traffic optimization, where "travelers" (data packages, vehicles...) might vie for priority while traversing through clogged infrastructures by outbidding others, and the value derived from the infrastructure's utilization is indirectly optimized through its owners' attempts to maximize its yield.

In these environments, existing YM algorithms and forecasting requirements are basically directly applicable. Practical implementation considerations would require that all involved subsystems (demand forecast, DSS, producer-consumer dialogue, resource assignment and actual execution of the assignment) are fully automated and seamlessly integrated.

Therefore, in most existing continuous environments that can be modeled in a way that lends itself to YM-type techniques, no interactive dialogue / customer decision making actually takes place. Heterogeneous customer queues waiting to be serviced by a set of alternative "resources" provide an illustrative example of these environments. They share a number of attributes with conventional YM environments. The cost of staffing a given set of resources might well be nearly independent of the customers actually served. Idle resources forego their contribution potential. Different customer requests might be of different values. If maximum acceptable waiting times are low, no "inventory" is allowed. However, the lack of the basic pricing based negotiation requires substantial adaptations in the YM approach, as discussed in the following sections.

4. Modeling continuous environments without interactive negotiation

The analysis of the application of YM-type techniques to continuous (and, by extension, hybrid) environments in which there is no interactive dialogue / customer decision making, exemplified by heterogeneous customer queues, has been carried out through a collaborative

project with a leading call centre operator. A six month in-depth, on-site analysis of the various types of call queues, the procedures used to assign calls to the various categories of operators, the supporting systems and procedures and the influence of the existing contracts/ Service Level Agreements provided the required framework (APQC, 2000).

These situations can be characterized as a moving-window assignment problem between customer requests waiting on a queue and alternative "treatments" for these requests.

Thus, the first element in the model is a sequentially numbered queue of customer requests "Req_r", where parameter "r" identifies the specific request. More generally, "Req_r" is a set of customer requests encompassing three ranges (not corresponding to contiguous "r" ranges): Queue (requests currently waiting in the queue), History (requests previously on the queue that have already departed through a "treatment") and Forecast (requests forecasted to arrive to the system, within the forecasting horizon). In call centers, these requests will correspond to customer calls.

The second element is a set of possible "treatments" to which a customer request or call can be assigned, "$Treat_t$". Parameter "t" identifies both the type of processing that the request will undergo and, where applicable, the specific resource that will carry out the processing (this can also be split in two layers, but that complicates the model). In a simple call center example, t=1, 2,...Op might denote the "Op" different operators that can handle that call; t=Op+1 might indicate the customer hanging up and t=Op+2 might mean that the customer is being transferred to an answering machine and asked to leave a message or call again later. In more complex settings, where the same operator can be instructed by the system to handle a request in different manners (e.g. try to understand the problem and agree to call the customer back vs. try to solve the request in just one call), t=1, 2,...Op might stand for the operator number "t" handling the call in the first manner, t=Op+1, Op+2,...2Op might denote the operator number "t-Op" handling the call in the second manner, t=2Op+1 might signify the customer hanging up, etc.

The endogenous decision that must be taken, and which the YM based DSS is supposed to support, is the assignment of requests "Req_r" from the "Queue" range to an element within the controllable subset of "$Treat_t$" (e.g, if t=Op+1 indicates the customer hanging up, that $Treat_t$ would be outside the controllable subset). In a call center setting, the decision is the assignment of a given customer call in the queue to a given operator, its transfer to an answering machine or the rejection of additional calls ("busy" signal). Decisions are taken sequentially, on a continuous basis (i.e., as time goes by, the system must decide whether to wait further or to make an assignment).

Exogenous events include: arrivals at the "Queue" range of Req_r, customer-controlled assignments (e.g. call termination) and changes in the "$Treat_t$" status (e.g. an operator completing the current call, or changing from available to unavailable status due to a pause)

For the DSS to be able to provide meaningful recommendations, information must be available on:

- Each Req_r

- Each $Treat_t$

- The "behavior" of the system

- An “objective function” linking how the various requests are processed to the objectives of the organization.

Each of these elements will now be analyzed using call-centers as a reference.

5. Application to the call center setting

This general model will now be applied to the representative call-center setting.

Req_r data

The information available for each request, Req_r, depends on the type of call queue being modeled and on the specific range within the request set to which it belongs (Forecast, Queue, History).

Two basic types of call queues were analyzed: for incoming calls and for transfer calls. Incoming calls include instances where the customer has not yet personally interacted with a human operator.

There are basically four data sources for incoming calls in the “Queue” range: the call time, the originating number, the dialed number, and the outcome of interactive voice response (IVR) dialogs.

The call time indicates the time at which the customer dialed in. When compared to the current time it provides the waiting time. Various “behaviors” (e.g. likelihood of hanging up) and most elements of the “objective function” are linked to the waiting time.

The originating number can be identified through Calling Line ID (CLI) services on conventional POTS lines or Automatic Number Identification (ANI) in telephony intelligent network services. Through Computer Telephone Integration (CTI), the Private Branch Exchange (PBX) sends that number (through an agreed upon protocol such as CSTA) to the system hosting the DSS (the same protocol will be used by the DSS, once an assignment decision is taken, to send that decision back to the PBX for implementation). Using that number, the identity of the caller can be looked up in the Customer Relationship Management system (CRM). In conventional CTI, this is done in order to, when that call is transferred to an operator that uses the CRM, automatically select the data for that customer on the operator’s screen (Screen Population or "Screen Pop"). However, for the purpose of the DSS described here, it also allows the DSS to access all customer data available at the CRM.

The Dialed Number Identification Service (DNIS) is applicable when several numbers are redirected to the same call center. For example, service numbers in various regions might be redirected to the same call center to achieve economies of scale, but, given the option, certain operators might be preferable for calls from certain regions. Another example more directly related to YM is the case where companies offer different numbers for customers from different classes (e.g. special numbers for preferential customers, or support numbers for customers with premium service contracts). Again these various calls might be redirected to the same call center to achieve load balancing, but they must be handled in a differentiated way (Durán, 2004).

As for the outcome of interactive voice response (IVR) dialogs, they may involve presenting the caller with menus of options and asking him to choose either through keystrokes (using

Dual-tone multi-frequency, DTMF) or through voice recognition. They may also involve identifying the user or even authenticating him through a password.

In Transfer queues, operators in one section (e.g. front desk) transfer customer calls to operators in another sections (e.g. subscriptions). Originating operators will normally have talked to the customer for a while before attempting to transfer the call, will have ascertained the customer's identity and will normally have fed the CRM with some additional data; at a minimum, procedures usually demand that operators enter a code in the CRM identifying the reason for the call (e.g. complaint, request for cancellation, subscription...). Time spent in the incoming call queue, in the conversation with the operator and in the transfer call queue should all be available. Thus, data available for Req_r transfer calls in the "Queue" range normally encompasses that of incoming calls plus additional, instantaneous information related to the call itself.

"History" calls have similar data sources to the transfer calls, except that operating procedures normally require operators to feed into the CRM some summarized, tabulated "closing" information upon completing a call.

As for "Forecast" calls, the source of (probabilistic) data is the forecasting module, that will normally draw on both a long-to-medium term forecast and a short-term forecast, derived from extrapolating the calls received immediately before.

$Treat_t$ data

Most $Treat_t$ (i.e., $Treat_t$ for most values of "t") correspond to individual operators (as discussed, if the DSS can instruct the same operator to handle a call in several different manners, there might be several $Treat_t$ for each operator). Relevant operator data encompasses reference and transactional data. Reference data is relatively stable and encompasses competences and services for which the operator has been trained. Transactional data is constantly updated and will normally come from the workforce scheduling and on-line monitoring system. It will include such data as whether the operator is currently unavailable, the call he is currently handling and since when, time remaining until the next scheduled pause or average workload in the last ten minutes.

Data for "treatments" which are note aimed at fulfilling the requests depends on the nature of that alternative approach. For example, if the treatment involves transferring the call to an answering machine, that treatment's transactional data will include the machine's current spare capacity. On the other hand, in a case explored in this call center project involving transfer calls, transferring operators had the option of, after the transfer waiting time exceeded a threshold, canceling the transfer attempt, informing the customer that he would be called back and opening an action request in the CRM, to be fulfilled by the destination operator whenever he becomes free; in this case, no system's capacity is involved.

System's behavior

This will include such reference data as average time to process a call or average time before a customer hangs up.

"Objective function" input data

A key input for the objective function is the Service Contract or Service Level Agreement (SLA) under which the call center operates. This is frequently a very specific and detailed

document itemizing how much will the call center receive for servicing each call, depending on a number of parameters such as the call type and the waiting time.

Therefore, in this call center setting, through the application of the appropriate technologies and operating procedures, the DSS is fed automatically in real time with all the relevant data to allow it to take continuous assignment decisions, which are then also implemented in real time (the PABX will route the call involved to the selected operator or automated service, will populate his screen with CRM data corresponding to that specific customer through CIT, and through the CRM will issue any additional instructions to the operator (such as the selected "manner", if applicable) .

6. Objective function

This DSS can be based on different approaches. For YM-type algorithms to be applicable, an essential element is a unified objective function which the algorithm attempts to maximize. Given the wide variety of inputs available for the DSS, that approach is only practical if several data elements can be combined into a smaller number of representative parameters.

The approach taken in this case was to synthesize all Req_r data, except call time/waiting time, into a single customer request classification parameter. That implies sorting all potential customer requests into a finite set of Customer Request Categories, $ReqCat_c$. This classification criterion must have three properties:

- "c" ≤ C, where C is the number of categories (as opposed to "r" in Req_r , which keeps on increasing)

- A call can be assigned to a specific category (that is, for each "r", c_r can be determined) based on the existing Req_r data (broader categories will be included for calls for which limited data exist, e.g. calls for which an existing CRM record has not been found

- That Customer Request Category summarizes the "revenue" implications of the available data for each request, as derived from the SLA

The objective function can then be modeled as a tridimensional matrix where:

Revenue (per call) = $R_{c,t,w}$

Where c represents the category code, t the treatment code and w the waiting time (w can be made discrete by breaking it down in intervals).

Completing each cell in this matrix implies specifying the "Revenue Contribution" that would result from processing a given type of request (e.g., from a customer whose CRM record indicates a "preferred customer") through a given "Treatment" (e.g. by servicing it through a highly qualified operator) after a given waiting time (e.g. after 10 seconds).

Once that matrix is completed, YM-type algorithms can be applied to establish the assignments that are more likely to increase the total sum of individual call revenues. They must take into account probabilistic considerations, e.g. while deciding whether to assign an available operator to a lower yield call that has been waiting longer (and is thus more likely to hang up), to a higher yield call with a shorter queuing time, or whether to keep he operator in reserve in case a yet higher yield call arrives.

Even though experience and continuous improvement can lead to efficient systems through trial-and-error, this type of systematic approach can sometimes reveal significant improvement opportunities. In this project, such an opportunity was identified through the analysis of transfer calls among sections. The established policy mandated that, once a customer call had been placed on hold for a time exceeding a threshold, the transferring operator should inform the customer that he would be called back and then open an action request in the CRM; that request would prompt the destination operator, when he becomes free, to contact the customer. Operators quickly discovered that, by waiting slightly longer than the threshold, their chances of obtaining a reply increased significantly, and thus gradually increased the real average waiting time. A structured "revenue" analysis showed that the established policy (and even more so the actual, deformed practice) was grossly inefficient when the system approached saturation. Since at peak times some calls had to be postponed anyway, there was no economic sense in having each call wait until nearly the maximum threshold (or even above it, with the actual practice). As soon as the DSS detected that the capacity was insufficient, the required number of calls could be postponed immediately, thus reducing not only their own waiting time but also that of the other calls.

7. Conclusions

Some continuous or hybrid environments, such as call centers, exhibit many of the traits of environments in which YM approaches are applicable. In most cases, however, some conceptual differences must be acknowledged, such as the inexistence of interactive negotiations with the customers. This requires engaging in some modeling activities, e.g. establishing an ad-hoc integrative objective function, before YM based DSSs can be productively applied. Furthermore, given the short decision times, the DSS must interact in closed loop with the environment. It is thus imperative that the forecasting subsystem is fed automatically in real time with the demand evolution (as through the automated-call-distribution integrated with the CRM in the call center case); the perishable asset availability schedule and the current reservations backlog must also be updated without human interventions (as through the workforce scheduling and on-line monitoring system in the call center setting, requiring operators to signal to the system its availability online); all these subsystems must be tightly integrated with the YM based DSS subsystem.

References

American Productivity & Quality Center (2000). *Call Center Operations. A guide for your journey to best-practice processes. APQC's passport to success series*.

Baker, T.K.; Collier, D. (2003). "The benefits of optimizing prices to manage demand in hotel revenue management systems". *Production and Operations Mgt*, 12(4):502-518.

Chiang, W.C; Chen, J.C.H.; Xu, X. (2007). "An overview of research on revenue management: current issues and future research". *Int. J. Revenue Management*, 1(1):97-128.

Duran, A. (2004). "Lean Potential of network-enabled remote service outsourcing: spatio-temporal decoupling and resource flexibility". *International Journal of Services Technology and Management*, 5(5/6):448-464.

Elliott, T.L. (2003). "Maximising revenue production while cutting costs: an airline industry mandate". *Journal of Revenue & Pricing Management*, 1:355–368.

Nair, S.K.; Bapna, P. (2001). “An application of yield management for Internet service Providers”. *Naval Research Logistics*, 48:348–362.

Pietraszek. W.E.; Ramchandran, A. (2006). *Using IT to boost call-center performance*. The McKinsey Quarterly, McKinsey & Co.