PREDICTION OF CUSTOMER MOVEMENTS IN LARGE TOURISM INDUSTRIES BY THE MEANS OF PROCESS MINING

Research paper

Brunk, Jens, European Research Center for Information Systems, University of Münster, Münster, Germany, jens.brunk@ercis.uni-muenster.de

Riehle, Dennis M., European Research Center for Information Systems, University of Münster, Münster, Germany, dennis.riehle@ercis.uni-muenster.de

Delfmann, Patrick, University of Koblenz-Landau, Koblenz, Germany, delfmann@uni-koblenz.de

Abstract

Customer movements in large tourism industries (such as public transport systems, attraction parks or ski resorts) can be understood as business processes. Their processes describe the flow of persons through the networked systems, while Information Systems log the different steps. The prediction of how large numbers of customers will behave in the near future is a complex and yet unsolved challenge. However, the possible business benefits of predictive analytics in the tourism industry are manifold. We propose to approach this task with the yet unexploited application of predictive process mining. In a prototypical use case, we work together with two major European ski resorts. We implement a predictive process mining algorithm towards the goal of predicting near future lift arrivals of skiers within the ski resort in real-time. Furthermore, we present the results of our prototypical implementation and draw conclusions for future research in the area.

Keywords: Process Mining, Prediction, Customer Movement, Tourism.

1 Motivation

In the past, advances in information and communication technology have revolutionized how we work, communicate, research or operate our business. The Internet and search engines in particular, have very visibly redefined how and where we get information on any topic in any life situation. Nowadays, new and less visible technological trends turn out to be just as transformative as the Internet. Big Data has been a hyped buzzword for quite a while. Many call the world we live in an Information Society or a data-driven world (Bryant et al., 2008).

In research, many different domains, including Information Systems (IS), tackle the opportunity to gain innovative and predictive insights from Big Data. Data driven analysis and prediction as a decision support system increasingly gets into the focus of researchers and practitioners, especially in regard to industry 4.0 (Lund et al., 2013). Goal of the analysis and prediction is always to support the respective entity in its business efforts and to optimize its processes (Shmueli and Koppius, 2011).

A core area of IS research is business process management (BPM) (Lee and Dale, 1998), which is of rising importance for both research and practice (e.g., van der Aalst, 2013; *Zukunftsthema Geschäftsprozessmanagement*, 2011). One industry that does not directly come to mind when talking about BPM is the tourism industry. However, the processes that underlie large tourism resorts (e.g., amusement parks or ski resorts) and public transport systems (e.g., subways or bike rental systems) represent typical characteristics of (logistical) processes. Such processes describe the flow of objects

(e.g., persons or resources) within networks through time and location (Delfmann et al., 2011). In many cities and countries, the tourism industry is an important, if not the most important, source of revenue (Unbehaun et al., 2008) contributing to regional growth and welfare (Paci and Marrocu, 2014). Therefore, predictive capabilities would provide the operators of such systems with a decisive benefit. They could actively steer customers, optimize process flows and save costs.

Ski resorts are a prime example of logistical processes. They consist of a network of lifts and slopes that customers flow through. We were involved in a data analysis project with two major European ski resorts. As it turns out, there exists no good prediction system for customer movement in large tourism industries, such as ski resorts. Typical approaches that are capable of evaluating big amounts of data towards the goal of analysis and prediction are based on machine learning techniques. Such approaches are often capable of predicting single instances of a business process (Metzger et al., 2015; e.g., De Weerdt et al., 2012). However, as soon as the behaviour of a large number of objects that flow through a business process needs to be predicted, these methods quickly reach their limits (Breuker et al., 2016). Although there are already predictive systems that can cope with vast amounts of data (like, e.g. (Castelli et al., 2016, 2017), who use geometric semantic genetic programming to predict average review scores of online store products), particularly predicting the way that flow objects will probably take in network systems based on their previous routes is currently not supported sufficiently.

We propose to solve this problem by the means of predictive process mining. Process mining is a "discipline providing comprehensive sets of tools to provide fact-based insights and to support process improvements" (van der Aalst, 2013), which deals with the analysis of event data being created during business process execution. In this domain, the evaluation and visualization of historic data has long been the primary focus of researchers and business applications. However, as in other domains, predictive analyses have become more popular and the novel area of predictive process mining receives more and more attention (van der Aalst, 2011). In today's digitized world, most business processes are supported through IS. These IS generate huge amounts of process data that can be exploited by process mining. Tourism resorts (such as ski resorts) also generate this data e.g. by the use of access systems.

In our work, we evaluate the applicability of predictive process mining to the prediction of customer movement in large tourism industries. To evaluate whether our approach is feasible, we implement it in a prototypical test case for the prediction of customer movements in ski resorts. Therefore, we answer the research question (RQ1), how can we understand lift usage of skiers in ski resorts as a business process and apply process mining to it? Our research goal is to develop a real-time prediction algorithm for future lift arrival of skiers. Through the realization of this goal we are able to evaluate, whether we can predict (near future) lift arrival of skiers through predictive process mining (RQ2).

2 Theoretical Background and Related Work

A core element of BPM is a process, which is "a completely closed, time-logical sequence of activities that are required for working on a process-oriented relevant business object" (Becker et al., 2011, p. 6). Consequently, a business process is "a special process that is directed by the business objectives of a company and by the business environment" (Becker et al., 2011). In contrast to a business process, a logistic process is a closed system of objects, where objects flow between a limited amount of positions (Delfmann et al., 2011). While business processes have historically been concerned about business process models and their notation, logistic processes have preferably been regarded under analytical aspects in terms of optimizing processes in the logistic industry. However, both business processes and logistic processes are part of BPM, which "is the discipline that combines knowledge from information technology and knowledge from management sciences and applies this to operational business processes" (van der Aalst, 2013).

In the discipline of process mining a process is discovered from event logs. Such events occur during the execution of business processes using IS and are logged in a log file. Each event represents a certain activity that was performed in a process and, hence, every event is of exactly one event type. Additionally, events belong to a process instance, i.e., multiple instances of the same process can occur in

parallel and each process instance may produce the same or a different sequence of events, depending on the activities performed in the process (van der Aalst, 2011). Therefore, event types, the timestamp when the event occurred and the process instance the event belongs to, are part of an event log file. In the first decade of the twenty-first century, the amount of literature on process mining has significantly increased (van der Aalst, 2013), making process mining one of the most important key concerns in BPM today. Not surprisingly, there is a great variety of process mining algorithms available (see de Weerdt et al. (De Weerdt et al., 2012) for an overview).

The case of predicting moving objects has been discussed in different fields of academia. For example, Chung et al. (2002) present an approach to mine temporal patterns for moving objects, which in their paper are mobile users to which location-based services are provided. Regarding the movement of pedestrians, models for prediction have been created from the 1970s on. For example, Haklay et al. (2001) provide pedestrian movement patterns and Asahara et al. (2011) provide a mixed Markov-chain model for prediction of pedestrian movement. Similarly, customer tracking at the point of sale has been successfully modelled with Markov chains as well (Krockel and Bodendorf, 2012). However, recent research shows that traditional approaches, like Markov chains, can be outperformed by neural networks and artificial intelligence (Gan and Lincoln University (Canterbury, 2005). Another method often used for simulation of pedestrian movement are cellular automata and multi-agent models (e.g., Tang et al., 2017; Weifeng et al., 2003). Such models provide a fine-grained simulation of pedestrians, as a grid of small cells is used to simulate bi-directional movement of individuals in a 2D or 3D environment. A benefit of cellular automata is that they can be rather easy visualized (e.g., Dijkstra et al., 2001). However, for application in our scenario of ski resorts, movements of individuals are only unidirectional (lifts are take uphill, slopes downhill) and a geographical location of individuals was not of interest. For these reasons, we have preferred means of process mining over cellular automata.

For large tourism industries, carrying out capacity management is both a challenge and a critical success factor, since it directly relates to their business performance (Cocossis and Mexa, 2004). Pullmann and Thompson (2002) have simulated a ski resort for capacity and demand management decisions. In a later publication, they suggest a combination of a conjoint analysis with a simulation model for profit maximisation in service networks, for which they again use the example of a ski resort (Pullman and Thompson, 2003). However, none of these approaches enables to predict the movement of individual customers. By regarding individual customers as process instances, i.e., as an object in a logistic process, we thus suggest to apply the tools and techniques of the process mining domain to customer movement prediction in large tourism industries.

3 The Case of Customer Movements in a Ski Resort

Ski resorts come in many different forms and sizes. At the time of writing 3.612 ski resorts exist in Europe alone, which run 15.927 ski lifts in total. Many resorts include only a handful of lifts. However, to stay ahead of the ever-increasing competition, the trend nowadays goes to the creation of super resorts. Currently the largest, Les 3 Vallées in France, spans a network of 143 ski lifts and offers up to 600 km of slopes. A variety of types of ski lifts exists, which transport skiers from the bottom of a mountain towards its peak. Most prominent are cable cars, chairlifts and draglifts. Once at the mountain station of a lift, a skier can choose between different slopes heading down the mountain to different locations and different valley stations of available lifts (Skiresort.info, 2017).

To control access to ski resorts, mountain railway operators make use of automated access systems. In modern ski resorts, these access systems represent mechanical turnstiles or modern boarding gates that make use of RFID technologies. To pass these access gates and to gain access to the ski resort, a customer purchases a ski ticket. The tickets are either equipped with a RFID chip or a bar code that the access system recognizes and then checks for permission to pass in the resorts central ticket system.

The majority of ski resorts set up access systems at each valley station of each lift. This ensures that only someone who owns a valid ski ticket can make use of the ski lifts and slopes. Some exceptions exist, e.g., in some glacier ski resorts where a very long cable ride is necessary to access the lifts necessary for skiing. Here, the resorts may use only one access system at the very beginning of the first

lift ride. In all other cases, the access systems log a very detailed event log of each skier passing through the valley station of a lift. Each ski ticket has a unique identifier. Therefore, one can trace a skier's movement through the ski resort by looking at live instance or past event log data.

We previously worked together with ski resort operators and a ski resort access system manufacturer. Based on our collaboration and their experience, we learned that two major matters of expense for ski resorts are their energy consumption and the repair of their lifts due to tear of the mechanics. Both factors directly relate to the speed at which a lift is running. In fact, lifts usually never run on full speed. Energy consumption as well as technical tear increase drastically the closer a lift gets to its upper speed limit. Therefore, it would be a great benefit, if, for instance, we would know beforehand that within the next 30 minutes only a handful of people arrive at a specific lift's valley station. Given that information, the lift operator can adequately lower the lift speed, and therefore costs, without causing an interruption of the guests skiing experience.

On the other hand, if we knew that soon much more skiers will arrive at our lift than our current velocity is capable of transporting, we could slightly increase the lifts velocity. Even though this incurs some costs to the operator, it has the capability to significantly improve the customer experience. This in turn could prove to be worth more than the added cost.

In general, the capability to predict and know how a skier will move within the ski resort in the near future enables the resort operator with many opportunities. One application could be to minimize lift queues via automatically adjusting lift speeds or optimizing skier distribution by actively steering them to less populated routes. Other possible opportunities include car like navigation systems with optimal routes or location based marketing activities of the resorts restaurants and après ski huts.

The aforementioned application areas of a near future prediction system of customer movement within ski resorts offers substantial economic potential. The lift access systems generate detailed event logs, which include cluster able information about the ski ticket user. However, so far no reliable prediction system exists, which lets the many resort operators use the available data.

4 Research Method

This paper follows the design-oriented information systems research paradigm according to Österle et al. (2011). As such, we construct an artefact, namely a concept for applying process mining technologies to the area of customer movement prediction. Thus, we follow a four-step research process, consisting of *analysis*, *design*, *evaluation* and *diffusion*.

Our research is motivated from the interest of large tourism industries, namely to predict customer movements for improving capacity management. Ski resorts are exemplary stakeholders. In chapter 1, we have identified and described the business problem and specified our research questions and goals (*analysis*). In chapter 5, we *design* our (algorithmic) solution to the specified case, which is *evaluated* for different scenarios in chapter 6. This paper itself presents the *diffusion* of our work to academia.

5 Prediction of Customer Movement

5.1 Preparation of Data

We received two datasets from two of the largest European ski resorts. Each data set includes all lift activity of every customer within the 2015/2016 winter season. That includes a time range from November 2015 to April 2016. In total, the two data sets together include over 30 million lift rides, and therefore event log entries.

The precondition to be able to use the data was, that it is only used for this research purpose and that we anonymize it in a way, such that no conclusion to the originating ski resorts is possible. For this reason, we do not use any ski resort names, lift names or details, which enable the reader to identify the ski resorts. Unfortunately, sometimes this also includes relevant information, e.g., exactly how many lifts exist in the ski resort or what type of lift a specific activity represents.

The first and smaller ski resort we from now on refer to as *Resort A*. The second one we refer to as *Resort B*. Both ski resorts are equipped with access control systems and have more than 20 lifts each. In both cases, these access systems registered more than 10 million lift rides during the 2015/2016 ski season.

Table 1 depicts the file structure of the CSV files with some fictive entries (due to anonymization). Each line of the file represents a single lift usage by a skier in the ski resort. The column *timestamp* refers to the moment in time that the skier passes through the access system.

timestamp	**datacarrier_id**	**validity_id**	**lift_id**	**lift_name**	**product_name**
2015-11-08 14:30:18	0123456789	123-456-78-90	123456	Lift 1	7 Days (Adult)
2015-11-08 14:30:19	1234567890	09-876-5432-1	123456	Lift 1	1 Day (Kid)
2015-12-03 08:04:33	9876543210	45-678-90-123	789012	Lift 8	Season (Kid)

Table 1. CSV Event Log File Structure

The *datacarrier_id* is the unique identifier of a RFID chip card. The card is used as a ski pass and can be reused multiple times by different customers over the season. The *validity_id* refers to a database entry that saves how long and for which parts of the ski resorts, a ski pass is valid. Therefore, the combination of the two columns *datacarrier_id* and *validity_id* serve as a unique identifier of a ski customer and his stay in the resort.

The columns *lift_id* and *lift_name* represent the id and name of the lift access system that registers the event. Here, it is important to distinguish between the lift and the lift access system. As the data evaluation showed, sometimes multiple *lift_id* and *lift_name* combinations refer to the same lift.

The last column *product_name* is a string description of the type of ski pass that the customer bought. This usually includes the amount of days the pass is valid, whether it is an adult, child or senior pass. But sometimes also additional information. Unfortunately, the manifestations of this string field are in no way standardized. Therefore, it includes a variety of entries, which sometimes refer to the same type of ski pass or only distinguish by one character that makes all the difference.

Instead of simply importing the whole CSV file and its column-structure into the database, we opted for a more advanced approach. The idea is, to parse each line of the CSV file through a PHP script and through this create a superior database structure. Figure 1 depicts the designed database structure.

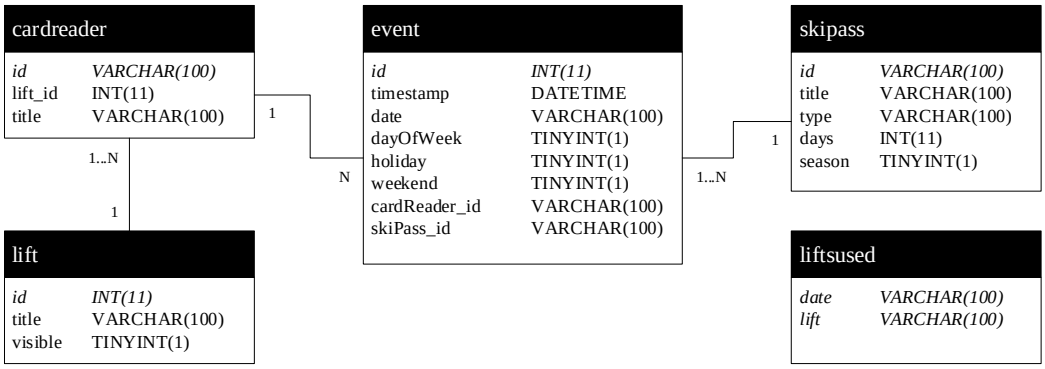


Figure 1. MySQL Database Structure

The first objective of this structure is to normalize the contents of the CSV exports. In the raw format, each line repeats all information about the access systems, RFID chips, ski tickets validity and ticket type with each event entry. The structure above splits this information into three tables: *event*, *skipass* and *cardreader*. A *cardreader* represents an access system at a lift.

The second objective of this step is to annotate and parse further information from the raw data. In the *event* table these include *date* (the timestamp without time information), *dayOfWeek* (Monday through Sunday), *holiday* (whether this day is a national holiday in the resorts country or neighboring countries) and *weekend* (whether it is a Saturday or Sunday). For the *skipass* table we executed a string

match on the *product_name* attribute to filter (where possible) the amount of *days* a ski pass was valid, whether it was valid for a whole *season* and the *type* of person group it can be assigned to (adult, child, youth, senior, ski instructor, etc.). Furthermore, we combined the *datacarrier_id* and *validity_id* to *id* within this table. These attributes combined are the unique identifier of a ski customer and his stay.

As previously mentioned, in a first analysis it turned out that in some cases multiple access systems (ids and names) refer to a single lift station. In addition, the naming of the access systems (cardreaders) is not very intuitive. To remedy this, we created a mapping together with the expertise of the access system manufacturer that unambiguously maps each *cardreader* entry to a lift in the ski resort. We then manually added all lifts into the lift table. The properties *id* and *title* are self-explanatory. We set the *visible* property to zero if we later want to exclude the lift from our calculations. For example, some access systems manage other kinds of attractions, which are no ski lifts. Alternatively, resort staff sometimes uses an access system only for test purposes. All normal lifts receive a visible value of one and are later included in the data set. Furthermore, we created the table *liftsused* that simply includes for each *date* value all the *lifts* that registered a lift ride on the respective day.

At this point, we export all the relevant lift activity (where visibility is set to 1) together with our additionally parsed information. This event log is now already cleansed from a number of entries that are not relevant to our research goals. We import the process data into the tool Disco, which is a process mining and automated process discovery tool. The tools strengths lie in its ability to work with big amounts of process data in a quick and responsive way (Rozinat et al., 2017).

Disco enables us to get a good overview over the process data. Insights include how skier activity spreads over the season or within selected days. In addition, which lifts skiers use most frequently and which ones they usually use in sequence. How many lifts an average user normally takes during a ski day, how long an average ski day lasts etcetera. Furthermore, Disco allows the user to define powerful filters on the data set. This feature is very helpful as it allows us to select only specific days or person groups to be included in the data set. In addition, we can export to different file formats, including the de-facto standard for process mining, called Extensible Event Stream (XES). XES files represent event logs. Therefore, they usually include standard components, such as *timestamp*, *case id*, *activity* and *resources*. When we imported our cleansed data into Disco, we used the *skipass id* and the *event date* as *case id*. This means, that each ski day of each customer represents one case (instantiation/execution) of the process of skiing through the resort. The *lift title* represents the *activity* and *event timestamp* represents the *timestamp* of the XES log. We also included the additionally parsed information as additional attributes, such that we can analyze and filter them within the tool.

5.2 Selection of a Process Mining Algorithm

In the literature, there are many different process-discovery techniques (see Tiwari et al. (2008) or de Weerst et al. (De Weerdt et al., 2012) for an overview). Such techniques are generally referred to as process mining and one of the first process mining algorithms was the α-algorithm, which was developed in the late 1990s (van der Aalst et al., 2004). The α-algorithm constructs Petri nets and can express concurrent behaviour in a limited way; for instance, the α-algorithm cannot express complex routing constructs, and neither can it express short loops. Therefore, various extensions have been developed like the $\alpha+$-algorithm (Li et al., 2007), which supports various kinds of loops, and the $\alpha++$-algorithm (Wen et al., 2007), which can also relate events which are not directly related in the event log.

While the general availability of process mining algorithms and techniques is huge, only very few of them consider probabilistic models, which are needed for prediction of customer movements. The aforementioned study by Weerdt et al. (De Weerdt et al., 2012) reviews twenty-six mining algorithms, of which only four were tagged as probabilistic. The only of the four approaches which is less than ten years old (Ferreira and Gillblad, 2009) is not suitable for prediction of process instances, since it concerns about assigning event logs to a process instance in case the process instance is unknown (i.e., the case of incomplete log files).

A recent publication in the renowned Management Information Systems Quarterly (MISQ) journal proposes a new process mining technique which utilizes probabilistic models (Breuker et al., 2016). In the publication, the RegPFA artefact is presented, a process mining algorithm which creates probabilistic models from event logs. Such models can be easily converted to process models, like Petri nets for instance, using tools like Petrify (Kishinevsky et al., 1997). Therefore, probabilistic models fulfil both the requirement for prediction of process instances and the requirement for mining a process model. For the presented case scenario, which is the prediction of customer movements in a ski resort, the RegPFA is a suitable process mining algorithm. Indeed, to the best of our knowledge, the RegPFA is currently the only process mining algorithm available, which fits our needs.

5.3 Implementation

For the implementation of our prediction design and the application of the RegPFA algorithm to the data of the ski resorts, we use the open source Process Mining Workbench (ProM). "ProM is an extensible framework that supports a wide variety of process mining techniques in the form of plug-ins" (Process Mining Group, 2017). ProM can read the above-mentioned XES file format and Breuker et al. (Breuker et al., 2016) previously implemented the RegPFA algorithm as a ProM plugin.

Before we can feed the previously cleansed data to ProM and the RegPFA, another data selection step is necessary. The RegPFA algorithm splits the event logs into a learning set and a control set. The learning set trains the probabilistic transition system, while the control set serves the purpose to test the model quality. This quality control on the one hand ensures that the model is not over fitted in regard to the learning data and on the other hand evaluates the quality of the identified transition system. However, in addition to that we need another control set. Because our goal is to predict future process outcomes, we need a second control set that we do not use to train the algorithm. This data set then simulates new live data from the ski resorts. At a specific time during the execution of the log, we will attempt to predict its future outcome. To evaluate the quality of our predictions we can check the actual process execution after the specified time of prediction to our predicted process outcome.

Aside from the control sets, we need to apply additional filters to the learning data. The result of the RegPFA algorithm is a state based transition system. This transition system replicates the typical flow of objects through the network of its states as well as possible. The question is what represents a typical ski day within our data set. To be able to predict a typical skiers behaviour as well as possible, we need to extract all strongly atypical data from our learning set. In the context of a ski resort, this can mean a variety of things. We analysed the data of the ski resorts regarding amount of skiers, point in time within the season, type of day, lifts open per day, event log case duration and number of activities per case. From our discussions with the domain experts and the analysis of the data, the amount of active skiers does not have a big influence on the route that a skier follows within the resort. The time a skier spends in the resort in a day (case duration) or the amount of lifts that he takes (number of activities) definitely have a major impact. However, these are exactly the behaviours that a probabilistic transition system is able to represent.

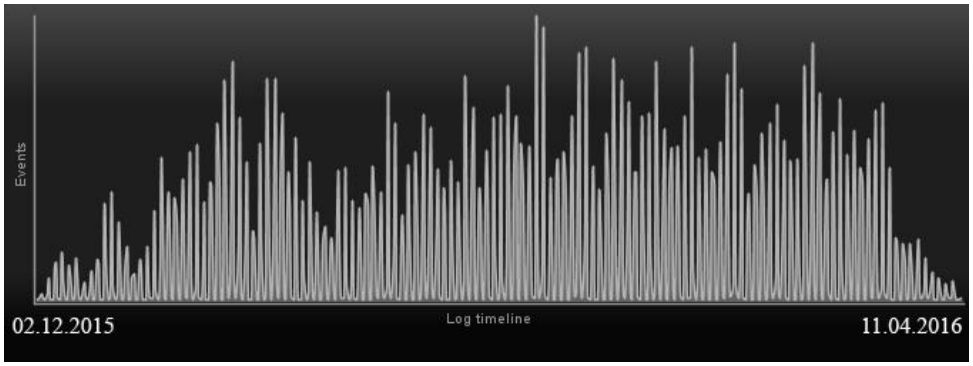


Figure 2. Events Over Time in Resort A

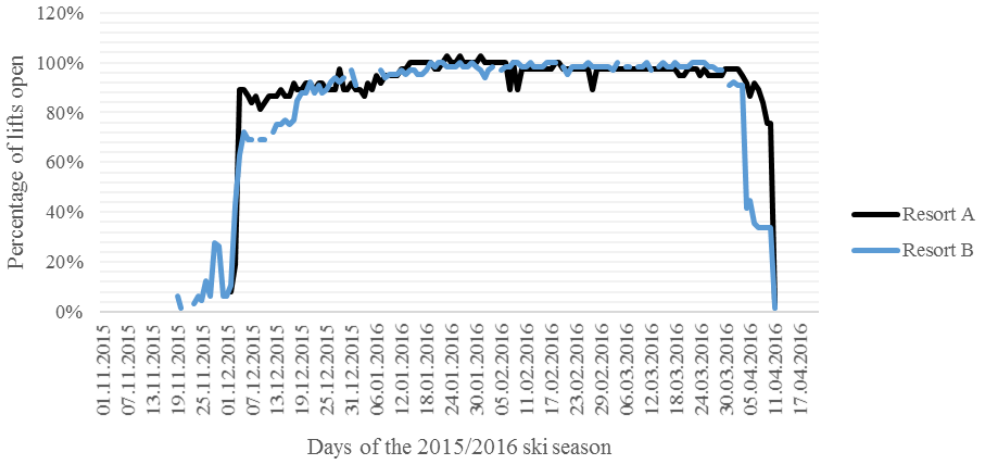


Figure 3. Lifts Open per Day in Resort A and B

Figure 2 depicts how the entirety of all log events (lift usage) spreads over the 2015/2016 ski season in *Resort A*. This directly correlates to the amount of skiers in the resort. As we expected, there is a significantly lower lift usage in the beginning and end of the season. Furthermore, throughout the main season, the figure shows regular spikes of lift usage. These are due to weekends or holidays in the respective countries. One reason for less skiers and less lift usage in the beginning and end might be the case that it is the off-season. However, it is the off-season for a good reason. Essential for the activity of skiing is a good amount of snow on the mountains. In Europe, snow usually falls from late autumn through Easter, depending on the location and height of the ski resort. In some places, the snow that fell persists longer than in others (e.g., due to amount of sunlight or topology). If there is not enough snow on a slope, as is more often the case in the beginning and end of the season, one cannot ski. If the slopes of a lift are not open for skiing, then it makes no sense to run the lift. Therefore, we analysed which lifts and how many of them were open for each day of the season. We recognized a lift as being open, if the according card readers registered customers passing through on the respective day (see Figure 3).

In fact, the amount of lifts open is a very important factor for the learning algorithm. How many and which lifts are running has a great influence on the possible routes a skier can take. Imagine a lift in the center of the ski resort that serves as a shuttle and connects different parts of the resort. If this lift is not running (due to amounts of snow, technical difficulties or too much wind), then the routes that the skiers take will greatly differ from a typical day where all lifts are active. Our goal in this prototypical implementation is to mine a process model that best describes the usual behaviour on a normal ski day.

Therefore, we limited our log data to those days on which most lifts were open. If we chose only the days where every lift was open, then the data sample would not be sufficiently big. On the other hand, with an increasing number of closed lifts the datasets get less comparable and the predictions that base on these datasets would be equivalently inferior. That is why we consider all the days where a maximum of two lifts were closed. In this case, the samples are satisfyingly big. For *Resort A* this applies to 73 days and for *Resort B* to 57 days of the entire season of 2015/2016. From this subset of days, we selected 5 days of *Resort A* and 7 days of *Resort B* for the control set of our prediction. The selected days represent different typical days of the week, holidays and weekends. We set the according filters on the date attribute in Disco and export the data sets as XES log files, which we then import into ProM.

The RegPFA algorithm needs one XES event log and configuration settings as input. The plugin automatically divides the input log into the learning and control set. For the configuration of the algorithm we refer to Breuker et al. (Breuker et al., 2016). In most cases, we applied the proposed standard settings, which make sure that the algorithm does not preliminarily terminate, but at the same time

finds as good results as possible. The most important configuration setting is the state range that the algorithm runs through. A number of tests on a smaller test data set show that the quality of the models continually increase with the number of states until it reaches its peak slightly above the number of lifts. After this point, the quality of the models, based on the included quality metrics, does not increase. However, the computation time continues to increase significantly.

State range	Runtime	Best solution properties	HIC value (quality criterion)
35 – 40	21h 47min 48s	States: 40 / Prior: 0,5 / Try: 3	2,91260665 x 10^7
41 – 45	25h 31min 52s	States: 45 / Prior: 0,3 / Try: 2	2,90314687 x 10^7
51 – 55	44h 31min 19s	States: 55 / Prior: 0,4 / Try: 5	2,879009655 x 10^7
56 – 60	69h 03min 25s	States: 60 / Prior: 0,4 / Try: 2	2.879689738 x 10^7
61 – 65	82h 44min 26s	States: 63 / Prior: 0,0 / Try: 4	2,872395868 x 10^7

Table 2. Learning Algorithm Execution Details for Resort A

A major challenge in the execution of the RegPFA on the large data set of the ski resorts was, that even though it supports multithreading, runtimes on local machines would take months to calculate for only one number of states. Therefore, we adapted the RegPFA plugin implementation so that you can run it in the experimental Command-Line-Interface (CLI) version of ProM. Given this adaption, we are able to execute the RegPFA algorithm and ProM on High Performance Computing (HPC) clusters that provide the needed computation power for the mining phase. In this instance, we were provided with access to *removed for blind review* from our universities central information management unit (*reference removed for blind review*). As ProM unfortunately does not support true parallelization, but only multithreading, the part of the cluster that we could use for our calculations is limited to six so-called Shared Memory Processing (SMP) nodes. These consists of 32 to 40 processor cores and 256 to 512 GB of RAM. As the HPC cluster is available to many researchers throughout the university, access to it is restricted through jobs and queues. The maximum run time for a single job on the longest running queue is therefore 160 hours. To prevent exceeding this maximum run time we need to split up the state range into smaller sub ranges. As the underlying Expectation-Maximization-algorithm follows an iterative structure, this is not a problem and we can later combine the results of the different subsets.

In case of *Resort* A, the execution of the learning algorithm for increasing state numbers continuously shows an improvement of the best solution found. The solution no longer improves from 63 states on (see Table 2). Theoretically, we cannot be sure that a better solution does not exist and that we only found a local optimum. However, due to the exponential runtime of higher state numbers, it is reasonable and only feasible for us to use the above found result as best possible outcome.

State range	Runtime	Best solution properties	HIC value (quality criterion)
65 – 69	153h 18min 24s	States: 69 / Prior: 0,1 / Try: 1	3,355552693 x 10^7
70 – 74	144h 57min 39s	States: 74 / Prior: 0,1 / Try: 4	3,346021414 x 10^7
75 – 79	Exceeded 160h	No final solution saved	
80 – 84	Exceeded 160h	No final solution saved	

Table 3. Learning Algorithm Execution Details for Resort B

In case of *Resort B*, which has significantly more lifts, this approach did not prove to result in an optimal result. Due to the higher number of states, even jobs with a small state range quickly reach the 160-hour limit (see Table 3). Therefore, for future steps we used the best result that was found with 74 states. According to the execution log of the algorithm, improvements in the HIC quality criterion were minimal and our expectation is that a local optimum would have been reached in one of the next iterations.

The above-mentioned results of the learning algorithm represent probabilistic transition systems. In these transition systems each state is connected to possibly every other state through directed edges. Each edge has an annotation that includes the activity name (lift name) that it represents and the probability of it occurring, given the current state. To understand how the prediction concept works, imagine the exemplary event log of Table 4 and the excerpt of a simplified probabilistic transition system of Figure 4.

timestamp	**lift name**
2016-01-23 08:30:13	Lift A
2016-01-23 08:59:45	Lift B
2016-01-23 09:15:06	Lift B

Table 4. Exemplary Event Log of a Skier

First, we follow each skiers' actions of the day from the moment he enters the ski resort (first registration at an access system) until the point in time where we want to make our prediction. To follow a skier means to evaluate in which state of our probabilistic model he could be in at the time of prediction. In this example, *s1* is the start state (quadrant 1 and 2). The first log event is *Lift A.* There are two edges of this activity leaving *s1* towards *s2* and *s4* (quadrant 3). The second log event is *Lift B.* For each state, one follows the available edges of this activity. In case of *s2,* we stay in *s2* and from *s4,* we move into *s6* and *s7* (quadrant 4). Our intention is to make a prediction at 9 o'clock. Therefore, for the prediction step we now consider only the three states of *s2, s6* and *s7.* Each possibility is weighted with the probability of reaching it via the edges. All outgoing edges represent the possible next lift the skier in question will take. The probability of each outgoing edge is multiplied with the probability of being in the state and the sum of all equal outgoing edges represent the overall probability of the skier using the according lift next.

We implemented a ProM plugin that executes this prediction based on a given probabilistic transition system, an event log and a timestamp at which the prediction is to be executed. In this context, the event log simulates the real-time application of our approach. As the event log, we use the previously extracted prediction control data set. With the plugin, we can predict the next lift activity of all skiers that are currently active in the ski resort.

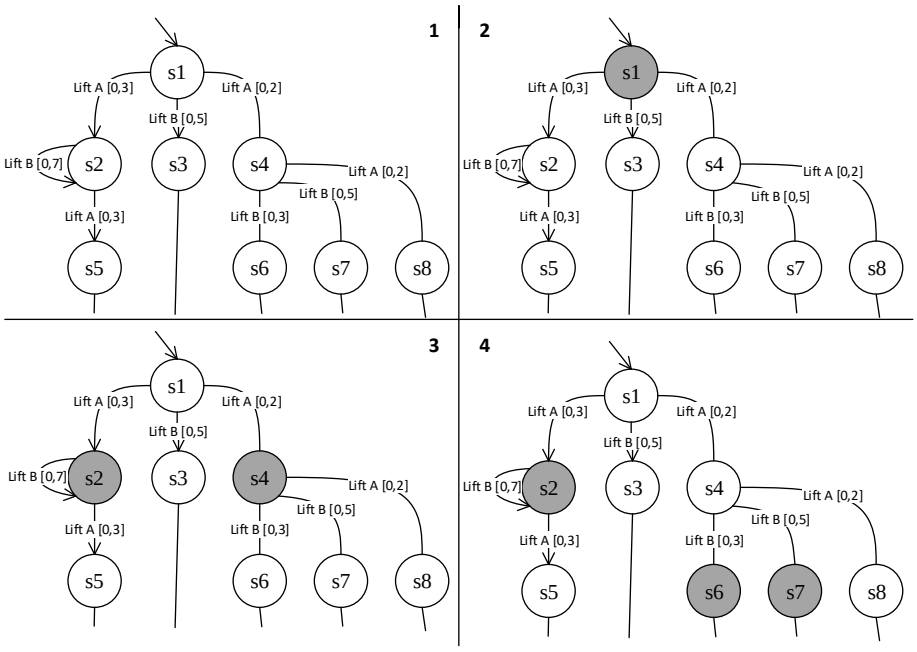


Figure 4. How to Follow a Skier through the Transition System

6 Results

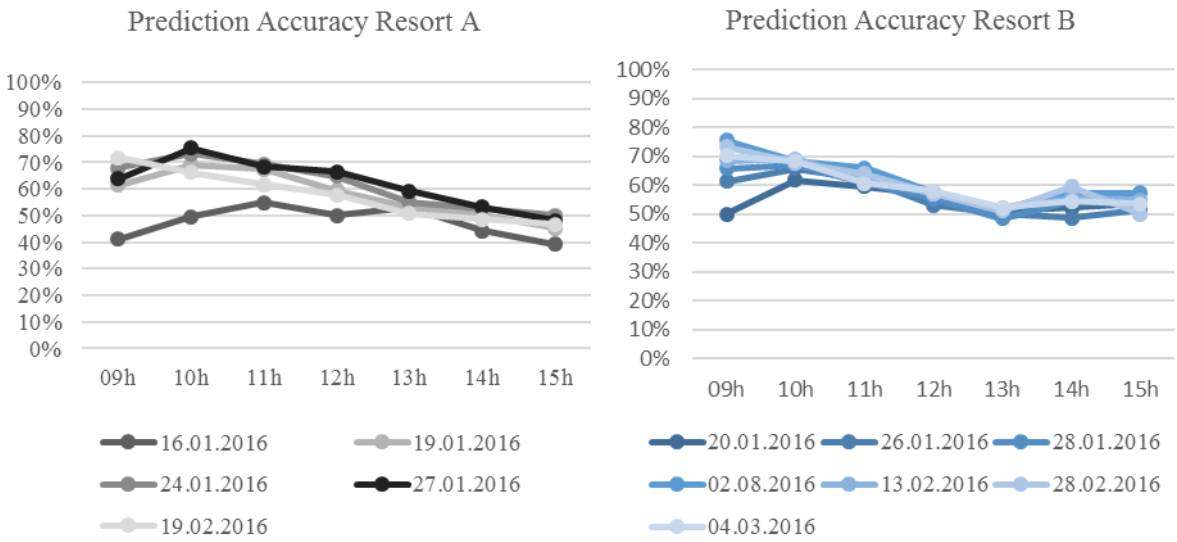
The process of the prediction follows the concept above. Each skier represents a trace in the event log. For each trace, we check whether the trace starts after the foreseen time of prediction or ends before it. We ignore these traces, as it means that at the time of the prediction they would not be present within the ski resort. For all other traces, we calculate the probability of which lift they choose next.

Table 5 depicts an excerpt from an exemplary prediction result of *Resort A* on *27.01.2016* at *10:00h*. Column one includes the anonymized lift name, column two the prediction of how many skiers will next use the respective lift and column three how many skiers actually took the lift based on the control set event log. The last column shows the calculated absolute deviation of the prediction from the log, which we use to calculate the mean percentage deviation of all lifts and therefore the overall prediction accuracy. For anonymization purposes, we unfortunately cannot include all lifts or the total number of traces in this table. We executed this process for all days of the control sets for both resorts. The execution of the prediction plugin can be execute in real time, as it only runs for a few milliseconds. For each day, we predicted the next lift activities at each full hour of the regular run times of the lifts (between 8:30h and 16:00h). The accuracy of these predictions is depicted in Figure 5.

The results show that the prediction accuracy ranges from 39% to 75%. The prediction constantly performs worse at the beginning and end of a ski day. Even more so towards the end. We used only one transition system to represent every skier's behaviour at all times. It makes sense that lift usage at the beginning and the end of the ski day deviate from the normal lift choice during the day. Figure 6 shows how the number of lift rides distributes during a typical ski day. The highest amount of lift rides take place between 10h and 11h. On the afternoon, there are significantly less lift rides. This can be an explanation for the constantly worse performance of the prediction during the afternoon. Less event log entries in the learning data means that the afternoon behaviour is represented less within the probabilistic model.

Next Activity by RegPFA: Resort A - 2016-01-27 10:00h			
Lift Name	**Prediction**	**Log**	**Absolute Deviation**
Lift 1	112.187959	58	54.18795888
Lift 2	24.8325525	39	14.16744747
Lift 3	26.0544763	9	17.05447631
Lift 4	128.410014	108	20.41001378
Lift 5	142.8201527	119	23.82015268
Lift 6	598.100393	622	23.89960703
Lift 7	129.238877	159	29.761123
Lift 8	100.4994843	70	30.49948431
Lift 9	417.5566805	384	33.5566805
Lift 10	118.2458837	83	35.24588373
…	…	…	…
Mean Percentage Deviation	**=** $\sum Abs.Deviation / \sum Log$		**= 0.24516427**
Overall Prediction Accuracy	**= 1 – Mean Percentage Deviation**		**= 0.754835727**

Table 5. Prediction Result of Next Activity – Resort A – 27.01.2016 at 10:00h



Even though the prediction accuracy is quite homogeneous, there are some outliers. For example, the prediction of *Resort A* day *16.01.2016* performs exceptionally bad. A closer look into the data reveals that on this specific day, there were only about half as many lift rides as on the other selected days. In addition, some major lifts only registered single digit lift rides. This could be a sign of extraordinary circumstances, e.g., a storm, heavy snowfall or fog.

The best prediction results lie in between 9h and 11h for both resorts. This is promising, because this is the time, in which the access systems log the most data and which therefore the calculated transition system best represents. Overall, the prediction accuracy is by far not good enough to apply it in practice. However, our goal is to show if and how predictive process mining can be used to predict customer movements in large tourism industries. This prototypical approach can serve as a basis for future research and development in the area.

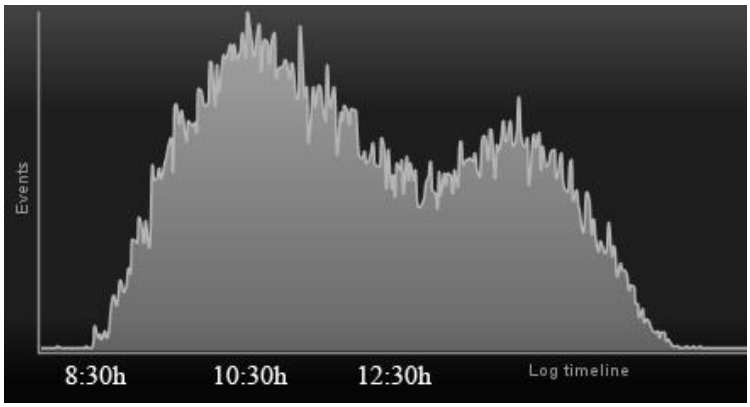


Figure 6. Distribution of Number of Events – Resort A 19.01.2016

7 Conclusion

We showed that one can interpret the lift usage of skiers as a business process and applied a process mining and prediction algorithm to it (RQ1). Therefore, we successfully parsed, cleansed and filtered the event log data of the lift access systems of two major European ski resorts. We applied the learning algorithm of the RegPFA on the cleansed data to calculate probabilistic transition systems. The transition systems, as state graphs, represent lift choice behaviour of customers of the ski resort throughout

the process (a ski day). We used these transition systems for real-time predictions of customer movements within the ski resorts. Towards our research goal, we developed a prediction concept, implemented it in a prototype and exemplarily executed it on the data set of the ski resorts. We used a control data set to simulate the live deployment of the system. In the end, we documented and presented the accuracy of the prediction prototype (RQ2).

Our results show that it is possible to apply predictive process mining to event log data of large tourism industries. In our example case, we applied it to the lift choice of customers of ski resorts. So far, the prediction accuracy is too poor to apply it in practice. However, there are reasons for this. We only considered the event log as input factor in this prototypical implementation. Many other important factors influence customer movement in large tourism industries. In the case of ski resorts, these include the amount of snow that lies on the slopes, the weather in temperature, wind, cloudiness/fog, rain or snowfall (the context of the process execution). In addition, we extracted information about the type of day and type of skier (age, person group, length of stay). Each groups of skiers or type of day might represent distinct movement patterns. Furthermore, the difference in movement decisions throughout the day is another factor that we intend to consider.

The above-mentioned limitations introduce areas of future research. An important next step is to include context information into the model creation and the prediction concept. One example is the lift status, such that we can consider all days of a season with all combinations of open/closed lifts in more specialized models. Furthermore, we are currently working on an improved version of RegPFA. It considers multivariate event types and thus can include context such as sensor data like temperature, wind, cloudiness/fog, rain or snowfall. We expect that this will greatly improve the prediction accuracy. Another area of future research is to enable true parallelization of the learning algorithm of the RegPFA algorithm. Then it is possible to calculate even more complex probabilistic models in significantly less time. To further increase the efficiency of the prediction algorithm, future research will also consider machine learning techniques that have already proven to work well on even huge datasets, such as genetic programming, like it was already proposed by (Castelli et al., 2016, 2017), together with their model quality measure (Castelli et al., 2011; Vanneschi et al., 2010). Similarly, if the exact location of skiers on downhill slopes can be tracked in the future, another look into the use of cellular automata and agent-models might be beneficial, as such technique could greatly help on calculation the utilization of slopes.

In this specific use case, our implementation lies the groundwork to support automatic or computer-aided rapid decision-making in large tourism industries. Using predictive process mining, for example a ski resort can automate the speed of their lifts to optimize customer flow and simultaneously save energy and repair costs. Furthermore, they could use the predictive power to steer customers through the resort to improve the overall distribution and customer satisfaction.

Acknowledgements

We thank the team of fluxicon, which stands behind the tool Disco that we applied in our research. They were kind enough to supply us with a full software licence free of charge for the purpose of our study.

References

van der Aalst, W.M.P. (2011), *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, Springer, Berlin/Heidelberg, Germany, available at: http://www.ncbi.nlm.nih.gov/pubmed/18487736.

van der Aalst, W.M.P. (2013), "Business Process Management: A Comprehensive Survey", *ISRN Software Engineering*, Vol. 2013, pp. 1–37.

van der Aalst, W.M.P., Weijters, A.J.M.M. and Maruster, L. (2004), "Workflow Mining: Discovering process models from event logs", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16.

Asahara, A., Maruyama, K., Sato, A. and Seto, K. (2011), "Pedestrian-movement prediction based on mixed Markov-chain model", *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, pp. 25–33.

Becker, J., Kugeler, M. and Rosemann, M. (2011), *Process Management - A Guide for the Design of Business Processes*, edited by Becker, J., Kugeler, M. and Rosemann, M., Springer, Berlin/Heidelberg, Germany.

Breuker, D., Matzner, M., Delfmann, P. and Becker, J. (2016), "Comprehensible Predictive Models for Business Processes", *MIS Quarterly*, Vol. 40 No. 4, pp. 1009-A9.

Bryant, R.E., Katz, R.H. and Lazowska, E.D. (2008), *Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society*, Computing Research Association.

Castelli, M., Manzoni, L., Silva, S. and Vanneschi, L. (2011), "A Quantitative Study of Learning and Generalization in Genetic Programming", in Silva, S., Foster, J.A., Nicolau, M., Machado, P. and Giacobini, M. (Eds.), *Genetic Programming*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 25–36.

Castelli, M., Manzoni, L., Vanneschi, L. and Popovič, A. (2017), "An expert system for extracting knowledge from customers' reviews: The case of Amazon.com, Inc.", *Expert Systems with Applications*, Vol. 84, pp. 117–126.

Castelli, M., Vanneschi, L., Manzoni, L. and Popovič, A. (2016), "Semantic genetic programming for fast and accurate data knowledge discovery", *Swarm and Evolutionary Computation*, Vol. 26, pp. 1–7.

Chung, J., Paek, O., Lee, J. and Ryu, K. (2002), "Temporal pattern mining of moving objects for location-based service", *Database and Expert Systems Applications*, Springer, pp. 331–340.

Cocossis, H. and Mexa, A. (2004), *The Challenge of Tourism Carrying Capacity Assessment - Theory and Practice*, Ashgate Publishing Ltd., Aldershot, UK.

De Weerdt, J., De Backer, M., Vanthienen, J. and Baesens, B. (2012), "A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs", *Information Systems*, Vol. 37 No. 7, pp. 654–676.

Delfmann, W., Dangelmaier, W., Günther, W., Klaus, P., Overmeyer, L., Rothengatter, W., Weber, J., et al. (2011), *Positionspapier Zum Grundverständnis Der Logistik Als Wissenschaftliche Disziplin*, Arbeitsgruppe des Wissenschaftlichen Beirats der Bundesvereinigung Logistik (BVL) e.V., Hamburg, pp. 262–274.

Dijkstra, J., Timmermans, H.J.P. and Jessurun, A.J. (2001), "A Multi-Agent Cellular Automata System for Visualising Simulated Pedestrian Activity", in Bandini, S. and Worsch, T. (Eds.), *Theory and Practical Issues on Cellular Automata*, Springer London, London, pp. 29–36.

Ferreira, D.R. and Gillblad, D. (2009), "Discovering Process Models from Unlabelled Event Logs.", *BPM*, Vol. 9, Springer, pp. 143–158.

Gan, C., Lincoln University (Canterbury, N.Z.. and Commerce Division. (2005), *Consumer Choice Prediction: Artificial Neural Networks versus Logistic Models*, Commerce Division, Lincoln University, Lincoln, N.Z.

Haklay, M., O'Sullivan, D., Thurstain-Goodwin, M. and Schelhorn, T. (2001), "'So Go Downtown': Simulating Pedestrian Movement in Town Centres", *Environment and Planning B: Planning and Design*, Vol. 28 No. 3, pp. 343–359.

Kishinevsky, M., Kondratyev, A., Lavagno, L. and Yakovlev, A. (1997), "Petrify: a tool for manipulating concurrent specifications and synthesis of asynchronous controllers".

Krockel, J. and Bodendorf, F. (2012), "Customer Tracking and Tracing Data as a Basis for Service Innovations at the Point of Sale", IEEE, pp. 691–696.

Lee, R.G. and Dale, B.G. (1998), "Business process management: a review and evaluation", *Business Process Management Journal*, Vol. 4 No. 3, pp. 214–225.

Li, J., Liu, D. and Yang, B. (2007), "Process Mining: Extending α-Algorithm to Mine Duplicate Tasks in Process Logs.", in Chang, K.C. (Ed.), *Advances in Web and Network Technologies, and Information Management*, Springer, Berlin/Heidelberg, Germany.

Lund, S., Manyika, J., Nyquist, S., Mendonca, L. and Ramaswamy, S. (2013), *Game Changers: Five Opportunities for US Growth and Renewal*, McKinsey Global Institute, available at: http://www.mckinsey.com/global-themes/americas/us-game-changers (accessed 27 November 2017).

Metzger, A., Leitner, P., Ivanović, D., Schmieders, E., Franklin, R., Carro, M., Dustdar, S., et al. (2015), "Comparing and Combining Predictive Business Process Monitoring Techniques", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 45 No. 2, pp. 276–290.

Österle, H., Becker, J., Frank, U., Hess, T., Karagiannis, D., Krcmar, H., Loos, P., et al. (2011), "Memorandum on design-oriented information systems research", *European Journal of Information Systems*, Vol. 20 No. 1, pp. 7–10.

Paci, R. and Marrocu, E. (2014), "Tourism and regional growth in Europe", *Papers in Regional Science*, Vol. 93, pp. S25–S50.

Process Mining Group. (2017), "start | ProM Tools", 22 November, available at: (accessed 22 November 2017).

Pullman, M.E. and Thompson, G. (2003), "Strategies for Integrating Capacity With Demand in Service Networks", *Journal of Service Research*, Vol. 5 No. 3, pp. 169–183.

Pullman, M.E. and Thompson, G.M. (2002), "Evaluating capacity-and demand-management decisions at a ski resort", *The Cornell Hotel and Restaurant Administration Quarterly*, Vol. 43 No. 6, pp. 25–36.

PwC. (2011), *Zukunftsthema Geschäftsprozessmanagement*, available at: http://www.pwc.de/de_DE/de/prozessoptimierung/assets/PwC-GPM-Studie.pdf.

Rozinat, A., Günther, C.W. and Niks, R. (2017), "Process Mining and Automated Process Discovery Software for Professionals - Fluxicon Disco.", 27 November, available at: https://fluxicon.com/disco/ (accessed 27 November 2017).

Shmueli, G. and Koppius, O.R. (2011), "Predictive Analytics in Information Systems Research", *MIS Q.*, Vol. 35 No. 3, pp. 553–572.

Skiresort.info. (2017), "Ski resort Les 3 Vallées – Val Thorens/Les Menuires/Méribel/Courchevel - Skiing Les 3 Vallées – Val Thorens/Les Menuires/Méribel/Courchevel", available at: http://www.skiresort.info/ski-resort/les-3-vallees-val-thorensles-menuiresmeribelcourchevel/ (accessed 27 November 2017).

Tang, T.-Q., Shao, Y.-X. and Chen, L. (2017), "Modeling pedestrian movement at the hall of high-speed railway station during the check-in process", *Physica A: Statistical Mechanics and Its Applications*, Vol. 467, pp. 157–166.

Tiwari, A., Turner, C.J. and Majeed, B. (2008), "A review of business process mining: state-of-the-art and future trends", *Business Process Management Journal*, Vol. 14 No. 1, pp. 5–22.

Unbehaun, W., Pröbstl, U. and Haider, W. (2008), "Trends in winter sport tourism: challenges for the future", *Tourism Review*, Vol. 63 No. 1, pp. 36–47.

Vanneschi, L., Castelli, M. and Silva, S. (2010), "Measuring bloat, overfitting and functional complexity in genetic programming", *... of the 12Th Annual Conference on ...*, pp. 877–884.

Weifeng, F., Lizhong, Y. and Weicheng, F. (2003), "Simulation of bi-direction pedestrian movement using a cellular automata model", *Physica A: Statistical Mechanics and Its Applications*, Vol. 321 No. 3–4, pp. 633–640.

Wen, L., van der Aalst, W.M., Wang, J. and Sun, J. (2007), "Mining process models with non-free-choice constructs", *Data Mining and Knowledge Discovery*, Vol. 15 No. 2, pp. 145–180.