Dynamic Forecast Model of Time Dependent Passenger Flows for Disruption Management

Evelien van der Hurk · Leo G. Kroon · Gábor Maróti · Peter Vervest

Received: Feb 17, 2012 / Accepted: date

Abstract The focus of recent research on disruption management in public transportation networks more and more shifts from operator-related aspects to improving the service quality. It is, however, not yet well understood how the passenger demand changes in a disrupted situation.

Forecasts of passenger demand traditionally focus on the long-term. Forecasts for disruption management on the contrary require detailed short-term predictions.

However, knowledge on passenger demand is generally lacking. Detailed information on passenger demand was not available until the introduction of smart card systems. Then again this data in its current form is not suitable for forecasting.

In this paper we propose a new algorithm that, based on the characteristics of the disruption, transforms the smart card data to suitable data for forecasting real-time passenger demand as required for disruption management. This is a new way of estimating demand for disruption management as well as a novel application of smart card data.
We present preliminary results for the Dutch passenger rail operator Netherlands Railways (NS).

**Keywords** Passengers · Disruption Management · Forecasts · Econometrics · Graphs

1 Introduction

Large numbers of people depend on public transport all over the world. Unfortunately disruptions occur regularly in these systems. Malfunctioning rolling stock or infrastructure, as well as computer system problems are just a few examples of the causes of disruptions. When a disruption occurs, complex rescheduling problems need to be solved quickly to adjust the logistic schedule to the new situation.

Detailed forecasts of passenger flows are needed as input for the current trend of disruption models focused on optimizing the passenger service level. However, traditionally demand forecasts are focused on the long term and therefore not accurate enough for disruption management, that requires real time passenger demand estimates including the origin, destination, time of departure and passenger type.

Until recently, detailed information on passengers’ journeys was not available. This has recently changed due to the introduction of smart card data. However, in its current form this data is not suitable for forecasting.

Our contribution consists of an online model for short term forecasts of passenger demand that not only converts smart card data into data suitable for demand forecasting, but also bases this conversion on a specific disruption. Hence the forecasts of passenger demand are tailored to the specific information needed for any disruption.

We will base our model on smart card data of Netherlands Railways (NS), the largest passenger railway transport provider in the Netherlands. We present preliminary results.

The remainder of this paper is organized as follows. In Section 2 we present the problem description of short term forecast modeling for disruption management. Section 3 contains a short literature overview of related research on forecasting, smart card data and disruption management. Our proposed algorithm for transforming smart data and prediction modeling is presented in Section 4. Some preliminary results based on smart card data of NS are presented in Section 5. We conclude this paper with a discussion and comments on future research in Section 6.

2 Problem description

Disruptions lead to complex logistic rescheduling problems because the timetable needs to be adjusted. Rolling stock needs to be rescheduled, new crew duties need to be generated, and all of this needs to happen quickly, while often not
knowing how long the disruption will last. At the same time, passengers are trying to figure out their best reaction to the current disruption. These two aspects depend on each other: the best action of the passenger depends on the rescheduled timetable, and the best rescheduling of the timetable and rolling stock depends on the reaction of passengers that results in the seat demand.

Optimizing the passenger service level requires detailed information on the passenger flows. Information on important transfers and routes requiring additional capacity is essential to make the right decisions on timetabling and rolling stock allocations. Alternative routes, available capacity, and incurred delay for specific journeys together determine the best travel advice given to passengers. Both the capacity replanning and the route advice to passengers reduce the delay of passengers and balance the demand and capacity.

A passenger’s best reaction to a disruption depends on his origin and destination. Furthermore, the reaction may depend on the type of passenger. For instance, a business card holder may have different alternatives available in comparison to a student. Also, a frequent traveler may have more knowledge on the train services and available alternatives than a passenger who travels only once a year. Finally, the time of the journey decides whether a passenger will be affected by a disruption, or not.

Consequently the origin, destination, time of departure and passenger type are all important information to estimate the changed demand for capacity, new connections that have become important to reduce delays, and the points in the system where passengers need to change their route due to the disruption management. Therefore, demand forecasts should include this information.

Until recently, detailed information on passengers’ journeys was not available. Information from train counts, ticket sales and surveys together enabled some rough estimate of OD flows, but data was never detailed enough to focus on short term forecasting of passenger demand. Only since the recent introduction of smart card ticketing, the time, ticket type, origin and destination of each journey is registered. Moreover, for each journey the card id is stored. As every passenger generally uses one smart card, this data enables analyzing journey patterns of individual passengers. In the Netherlands, smart card data is available to the Public Transport Operator with one day delay. Hence it is a great source for short term forecasts.

Smart card data in its current form is not directly suitable for forecasting. We need to transform this vast amount of data, while assuring that the essential geographical and time information is still preserved. We propose an algorithm that, given a specific disruption, adapts the way smart card data is aggregated into suitable data for short term forecasting. The transformation takes into account the journeys’ temporal and geographical information as well as the ticket type and the travel pattern of the passenger.

Forecasts of passenger demand include information on the origin and destination of the destination. Forecasts are estimating demand for capacity before passengers react to the disruption. The predictions are independent of the path choice of passengers. Therefore they provide a good basis to estimate additional demand for capacity given a specific new timetable.
We define an online prediction model that forecasts real-time passenger flows based on a specific disruption, by first transferring data and then using this specific data for forecasting. To our knowledge this is the first research to focus on short term forecasts of passenger demand for disruption management. Hence our contribution is not so much focused on finding the most accurate predictions, but mostly addresses what information on passenger flows is needed for disruption management and how to obtain this from smart card data.

We address the following research question:

*How can we deduce information for route advice and rolling stock rescheduling from smart card data?*

Second research question:

*What forecasting models perform better: separate models for specific passenger groups, or one model on the aggregated number of passengers?*

2.1 Dutch Smart Card Data

As smart card data is important in this paper, we provide some background information.

Smart card systems are a way of ticketing. In the Dutch system, which is similar to the systems in Tokyo, Seoul, Singapore and the London subway, passengers need to tap their card on entrance and exit of each journey. For railway transport in The Netherlands, the devices for checking in and out are at the stations. Consequently the data generated by smart cards stores the start location, end location, start time and end time of each journey per card - but not the specific route of a passenger. Each card is uniquely defined and linked to a specific product type. Data on passengers’ journeys was never before available in this detail.

In the near future, all passengers will have to check in and out for each train journey. Currently, smart cards are still in the introduction phase. They are not mandatory (except for student subscriptions). Alternatives in the form of paper tickets and e-tickets are still in use. The most frequent travelers, the passengers with a month or year subscription for unlimited travel on either the full network or part of the network, do not use smart cards currently. Hence the current data set is limited: it does not register all journeys, nor is it a random sample from the passenger population.

The full introduction of smart cards is planned for the start of 2013. Therefore in the very near future, smart card data will contain information on all journeys. Moreover, similar systems are available around the world: from Tokyo to London. This paper is therefore mostly about how to use this kind of detailed data to come to accurate predictions of passenger flows: a method that is applicable to any set of data similar to smart card data.

We use the smart card data of the NS as a test set for our methodology, containing a significant amount of total journeys. Because the regular passen-
gers, the passengers that are probably most easy to predict, are the ones that lack from our data set, we expect the resulting forecast accuracy to be at least as good for the full set of journeys as it is for the current sample.

3 Literature Review

In this Section we give an overview of related research to short term forecasts of passenger flows for disruption management. We start with literature on demand forecasting for transport, follow by a short overview of research on the application of smart card data and concluding with some key references for disruption management.

3.1 Forecasting OD flows and demand for transport

In this section we focus on research on demand forecasting for public transport. Before smart card data, these models usually relied on panel data or aggregated data, and focused on long term predictions of demand and elasticity’s.

Some research focusses on long term predictions of demand, like Gaudry (1975) who uses aggregated data on the demand for public transport in a specific urban area, to forecast future demand based on the price of public transport, price of alternatives to public transport and the demographics of the population, like the income distribution of its inhabitants. Wardman (2006) specifically focusses on forecasting railway demand in the long run, using similar variables as the model of Gaudry (1975).

Others focus on the elasticity of demand. Hsiao and Hansen (2011), for instance, focus on predicting demand for air passengers, combining in their model predictions for demand generation and assignment, and base their model on panel data. They focus on predicting the sensitivity of demand to both time and price of the journey. Rolle (1997) also forecasts the elasticity of demand to price, specifically for railway demand, taking into account the different kind of services offered. Batley et al (2011) focusses explicitly on the long term effects of lateness of trains on railway demand. He compares a market level model with an individual level model based on panel data, and comes to the conclusion that the effects on the aggregated level in terms of elasticity is less severe than the individual level models based on panel data suggest.

From the area of complexity research, González et al (2008) show that human mobility patterns are very well predictable, based on a US experiment using cell phone location data. The AURORA\(^1\) project that focuses within NS on predicting numbers of passengers per train, is very successful in their approach based on several data sources, that since recently also include smart card data.

\(^1\) AURORA is the name of a series of projects focused on long term predictions of the number of passengers per train.
We are interested in short term predictions of passenger flows as opposed to the long term predictions and focus on elasticity’s that is prevalent in most of the above literature, and in the specific focus on OD flow prediction. The success of long term predictions for similar kinds of flows, and the different approaches from both macro and micro models, are an indication of the viability of this approach.

3.2 Research on Smart card data

Literature on smart card data and applications of smart card data mostly stems from the start of the current millennium. Blythe (2004) was one of the first to give a functional overview of the introduction of smart card systems for ticketing in public transport. The recent article by Pelletier et al (2011) provides a review of literature focused on analyzing smart card data. They divide the literature into the categories strategic level, tactical level, operational level and commercialization. We use a different categorization of those articles on smart card data that are related to this paper, and divide them into the sections passenger behavior and route choice and OD. We refer the interested reader to Pelletier et al (2011) for a broader review of smart card data.

Analysis of passenger behavior

One of the uses of smart card data is for analyzing passenger travel behavior. Bagchi and White (2005) are one of the first to use smart card data. They use smart card data of a UK bus company, where they analyze the passenger population of that company. Similar is the work of Park and Kim (2008) who analyze usage and mode choice of passengers based in smart card data of the Seoul public transport system.

Morency et al (2007) and Agard et al (2006) extend these analyses by using data mining techniques to deduce travel behavior for different groups of passengers. They show that clustering techniques can reveal patterns of groups of passengers that are valuable for the analysis of transport usage for a PTO. They focus on the analysis of past performance and behavior of passengers. Also, in their data sets the destination of the journey is not included.

Zureiqat (2008) focuses on the prediction of passenger travel behavior for revenue management. He focusses on the prediction of product type choice together with the number of journeys per passenger. He develops a model for forecasting the effect of product and price changes of a PTO.

Route choice and ODs

Second to the analysis of passenger travel behavior, few articles exist on estimating route choice and ODs from smart card data. These estimations are in general focused on in hindsight constructing route choices and ODs of passengers, as the destination is not registered in every source of smart card data.
Seaborn et al (2009) analyze smart card data of the London system, investigating transfer times between different legs and modes of transportation. Kusakabe et al (2010) focus on linking check ins and outs to trains for the Tokyo system. In this system they focus on one line having different services. Based on the check ins and check outs, they deduce the route choice as the (combination of) services the passenger has chosen for his journey. The algorithm they propose is very similar to the algorithm used in practice by NS for linking journeys to routes.

We also mention here the work of Zhao (2004) and Gordillo (2006) because they focus on OD matrix estimation, although their problem setting is different from the one studied in this paper. In their paper, they do not aim to forecast the OD flows, but focus on deducing these from historic data, as the end locations of trips are not registered. From the estimated OD matrix, secondly the routes of the passengers are inferred. The article focuses on the London bus system in which, different from the London subway system, just a single check in is required.

Summary smart card data research

Current literature on the analysis of smart card data shows that it is a valuable source of information to deduce and qualify passenger travel behavior. However, to our knowledge, none of the research has focused on forecasting passenger behavior or OD flows from the smart card data.

3.3 Disruption Management

Caprara et al (2007) give an excellent overview on Operations Research (OR) in passenger railways. They present many operations research models used for line planning, rolling stock circulation, crew planning and shunting, among others.

The focus on passengers in disruption management is rather recent, of which good examples can be found in Dollevoet et al (2010), Nielsen (2011) and Veelenturf (2010). In their research on disruption management they take a passenger oriented approach. Minimizing passenger delay is the objective of their research. They include the reaction of passengers to a disruption. However, they assume initial OD flows given, while in practice this is in general not true. Therefore the results of this paper would be valuable input for their research.

4 Methodology

Our methodology consists of two steps. First, we convert smart card data into several time series, second we use these time series for prediction modeling. Figure 1 shows the process of first converting smart card data over the dimensions of geographical information (ODs, Section 4.1), time information (start
and end time of the journey, Section 4.2) and the passenger type (product type and journey pattern, Section 4.3). For each of these dimensions we need to decide what information to aggregate, and what information to include in the time series - either as variable or as separate time series.

The second step is to forecast passenger flows based on this data. Our aim is to forecast passenger flows independent on route choice and also before their reaction to the disruption. This demand forecast then provides all information to estimate demand for additional capacity, to indicate important transfers and to know the decision points in the network where passengers need to change their path due to the disruption.

As shown in Figure 1 there are two actions before we arrive from transformed smart card data to demand forecasts: first we need to select a model type, and secondly we need to select the variables that add prediction power to the forecasting model. Section 4.4 describes the approach we plan to take for forecasting, but we cannot present results yet.

\begin{figure}[h]
    \centering
    \includegraphics[width=\textwidth]{figure1.png}
    \caption{First we need to convert smart card data to data suitable to estimate forecasting models.}
    \label{fig:figure1}
\end{figure}

4.1 Clustering of relevant passenger flows

The Dutch railway network contains almost 400 stations, resulting in a set just short of 160,000 unique OD pairs. Per day, NS provides approximately 1.1 million journeys. If all unique OD pairs had an equal number of passengers traveling along them, there would be less than 7 passengers per specific journey per day. However, in reality a small set of ODs has the majority of passengers, and therefore a large set of journeys will have few passengers traveling along them. Consequently, estimating flows for all ODs not only leads to an enormous number of forecasts, it also may prove to be hard to predict flows accurately for many ODs because so few people travel between them.

We would like to cluster ODs based on their geographical location but not lose important information. For disruption management, passengers that need to change their route are important. Before that decision point passengers will not change demand for capacity, or require additional information.
But from the point on where the routes start to differ, they change demand for capacity. Also, before reaching that point of change they need route information to make the best decision. Hence these points in the network where passengers need to change their route and rethink their decision; those contain the important geographical information. We will refer to them as decision points.

For clustering destinations the approach is similar. Once a passenger is on his or her normal path to the destination, the demand for seats and demand for information may be as expected: the passenger is back on the path he or she always planned to be at. Although the point where the detour and the original path come together is more of an ‘anti decision point’ for ease of notation we will also refer to these as decision points.

We propose an algorithm for finding these decision points in the network and cluster ODs based on these decision points. This results in a kind of dictionary, where every OD is linked to two specific decision points, but this pair of decision points may have multiple ODs referring to them. Before going into the details of the algorithm, we first present an example for illustration.

**Relevant OD flows and Decision Points**

In the transformation process of smart card data to forecasting data, we cluster ODs to keep the essential geographical information. The following example is illustrated in Figure 2. We illustrate the clustering algorithm by showing how OD flow $MG$ is clustered to $JG$, given a disruption on $GH$ that breaks the connection between the two stations. As we focus on finding major decision points, we will focus only on major stations in the network. Regional stations can in a post processing step easily be assigned to the closest decision point.

First, based on the full network we find the shortest path from $M$ to $G$, and the shortest path from $M$ to $G$ through the disruption, $GH$. Figure 2a shows the network and the connection. The vertices are the stations and the edges are the connections between the stations. Every specific type of train service has a separate edge. If we were to focus on $LM$, the algorithm would terminate here: clearly passengers traveling from $L$ to $M$ will not be affected by a disruption at $GH$ and therefore they are excluded from the demand forecasts. However for $MG$ the shortest path is equal to the shortest path through the disruption, hence we proceed with the clustering.

First we reduce the network to contain only the fastest connections between stations, and only look at major stations in the network. Transfers are excluded as because of the disruption we might want to or have to change these. The network then looks like Figure 2b.

Secondly, we find the decision nodes for $MG$. First we look at the shortest $MG$ path in the normal network and compare this to the shortest $MG$ path avoiding the disruption. Both paths are drawn in Figure 2c. The first is $MKJHG$ and the detour path is $MKJIABG$. Comparing these paths, we see that they are equal until station $J$. At this station, a decision needs to be made. To check if this is the first station where a decision needs to be made,
(a) Find relevant OD flows through shortest path computations

(b) Reduce network to main stations and quickest connections for cluster computation

(c) Example of assigning an OD flow to a cluster

Fig. 2: A general overview of the steps taken to find relevant OD flows and cluster these based on the rerouting options. Illustrated by an example of a railway network.

We delete node J from the disrupted network as well. Clearly, no other paths are possible from M to G. Therefore J is the origin decision node of path MG.

Then we shift our attention to the destination node, G. Looking again at the two paths MKJHG and the detour path MKJ1ABG we see that there are no other stations coinciding between J and G. Therefore G is the destination decision node of MG.

We register MG in the dictionary as belonging to JG and move to the next OD pair.
**Algorithm for Clustering OD flows**

Here we give a more detailed description of the algorithm for clustering ODs. We start with choosing a threshold $\alpha$. Any journey no more than $\alpha$ time units longer than the shortest path is considered to be a reasonable path in the undisrupted network. In the computation of reasonable paths in the undisrupted situation, we will include transfer times. We are interested in normal planned journeys, and transfer times can make the difference between a path being reasonable or not.

Choose thresholds $\beta$ and $\gamma$. A reasonable detour exists when the shortest path in a disrupted network is less than $\beta$ time units longer than the shortest path in the undisrupted network. Furthermore, a reasonable alternative detour exists when a path is less than $\gamma$ time units longer than the fastest detour. We use two parameter settings here because passengers will accept a slightly longer travel time to get to their destination, but at a certain additional delay their journey may become pointless. Although $\beta$ may be personal, we will assign a global value to this parameter. Although passengers may be willing to travel $\beta$ time units longer if this is the only way, not all paths taking this long may be considered. The parameter $\gamma$ indicates that the accepted delay is restricted relative to the shortest detour.

In the computation of paths in the disrupted situation, we exclude transfer times. In this situation, we want to find the important transfers on rerouting paths. Enabling or improving such a transfer can significantly reduce delays of passengers, especially when the difference is small. Just missing a transfer will lead to a large delay and hence less attractive rerouting options, while these options may be the easiest to speed up by delaying the connecting train just a little bit. Including transfer times may not indicate these paths as reasonable or short, while they are in fact very interesting. Excluding transfer times solves this problem.

We consider an OD to be relevant whenever there is a reasonable path that goes from O to D through the disrupted area. By comparing the length of the shortest path from O to D and the shortest path from O to D through the disrupted part of the track we can easily decide whether a specific OD is a relevant flow.

Secondly, we need to find how these ODs can be clustered into a more condensed set of ODs. Outside the disrupted area, we assume all train services run as planned. Therefore the origin and destination only becomes relevant when options for rerouting occur. Hence we are interested in those points in the network where a route decision needs to be made. Note that we aim to cluster the ODs by making a dictionary matching original ODs to clustered groups $\tilde{O} \tilde{D}$. Consequently every OD pair may be assigned to one $\tilde{O} \tilde{D}$, where OD=$\tilde{O} \tilde{D}$ is allowed. We choose these $\tilde{O} \tilde{D}$ points as the latest point in the route where the first route decision needs to be made (O to $\tilde{O}$) and as the latest point at which different routing options first join again ($\tilde{D}$ to D). We only consider reasonable paths, though in case of a disruption we choose a different threshold $\beta, \alpha \leq \beta$, to find reasonable paths.
We would like to know the decision points important to the set of all reasonable paths - which would lead to finding the \(k\)-shortest paths. However, as we will assign a single OD pair to just one new \(\tilde{\text{O}}\text{D}\) combination, we only need to find the decision points and not the path itself.

We define:

- \(\text{O}'\text{D}'\): a specific OD pair from the set of all ODs
- \(\tilde{\text{O}}\tilde{\text{D}}\): a pair of two stations that possibly represents multiple OD flows
- \(d(\text{O}'\text{D}', \text{N}', \text{X})\): the journey time from \(\text{O}'\) to \(\text{D}'\) in network \(\text{N}'\), either through path \(\text{X}\) or through a shortest path if no \(\text{X}\) is provided
- \(p(\text{O}'\text{D}', \text{N}', \text{X})\): the order of stations in the path from \(\text{O}'\) to \(\text{D}'\) in network \(\text{N}'\), either through path \(\text{X}\) or shortest path if no \(\text{X}\) provided
- \(k\): a disruption either blocking or delaying a subset of tracks
- \(N\): full network containing all links between all stations
- \(N_k\): network as changed by disruption \(k\)

We propose the following algorithm for clustering ODs.

**For all ODs, given a disruption \(k\) in a network \(N\):**

1. Take an \(\text{O}'\text{D}'\) from the set of ODs. Set \(N\) and \(N_k\) to initial values.
2. Set path \(A\) as \(p(\text{O}'\text{D}', N)\) and path \(B\) as the shortest path in the network going through any part of \(p(\text{O}'\text{D}', N)\)
   - If \(d(\text{O}'\text{D}', \text{N}, B) > d(\text{O}'\text{D}', \text{N}, A) + \alpha\), go to step 1
   - Else go to step 3
3. Set path \(C\) as \(p(\text{O}'\text{D}', \text{N}_k)\). Set the Minimal Detour Time \(S\) equal to \(d(\text{O}'\text{D}', \text{N}_k)\)
   - If \(S > d(\text{O}'\text{D}', \text{N}, A) + \beta\), then add \(\text{O}'\text{D}'\) to the set of journeys for which no reasonable detours exist. Go to step 1.
   - Else go to step 4
4. Compare \(p(\text{O}'\text{D}', \text{N}, B)\) and \(p(\text{O}'\text{D}', \text{N}_k, C)\). Set \(\tilde{\text{O}}\) equal to the first point where the paths start to differ and \(\tilde{\text{D}}\) to the latest point at which the paths are first joined.
   - Else if: \(\text{O}'\text{D}' = \tilde{\text{O}}\tilde{\text{D}}\), add \(\text{O}'\text{D}'\) directly to the dictionary. Go to step 3.
   - Else go to step 5
5. Delete \(\tilde{\text{O}}\) from the network \(N_k\).
6. Set path \(E\) as \(p(\text{O}'\text{D}', N_k)\).
   - If \(d(\text{O}'\text{D}', N_k, E) \leq S + \gamma\) path \(E\) is a reasonable path, set path \(C = \text{path } E\), repeat step 4
   - Else go to step 7
7. Fix \(\text{O}' = \tilde{\text{O}}\). Restore \(N_k\) to initial values, then Delete \(\tilde{\text{D}}\) from the network \(N_k\).
8. Set path \(F\) as \(p(\text{O}'\text{D}', N_k)\)
- If $d(O'D', N_k, F) \leq S + \gamma$ this new path is a reasonable path, set path $C = \text{path } F$, and repeat step 4
- Else Add to Dictionary: $O'D' = \tilde{O}\tilde{D}$. Go to step 3

The algorithm requires only shortest paths requests. Given the nature of the network, these computations will be efficient. The algorithm uses a clear but flexible concept of reasonable paths. Furthermore, as it deletes a node in the iterations per specific $O'D'$ we know that the algorithm will terminate. Also, it will assign one and only one value of $\tilde{O}\tilde{D}$ to a specific $O'D'$ and hence the presented solution is unique.

4.2 Time

Start and end time of each journey are stored in a precision of seconds. Splitting up data to this grid not only makes forecasting hard because of the lack of numbers, forecasts per second could also easily drown planners in information. Moreover, a second of time can make the difference between catching a train - or not. Because we are interested in rough estimates of passengers for disruption management, we choose to aggregate this time information in manageable blocks.

*Generating flexible time intervals*

Disruptions that occur and lead to complex replanning problems will need several minutes in the least for the decision and implementation process. Therefore, forecasts of passenger flows can be over a longer time interval. However, the volume of passengers does strongly depend on the time of day. By taking a fixed time interval with a flexible start time, we circumvent the problem of too much detail but do capture the most recent arrival information of passengers.

Within this paper, we will look at 30 minutes, 1 hour and 2 hour intervals with a flexible start time. This is motivated by our test data set of NS smart card data. Because NS runs a cyclical timetable with cycles of one hour, with 4 trains an hour between major stations and 2 trains an hour for most other stations, these intervals come naturally. However, our approach is suitable for any time interval smaller or larger than the proposed intervals here.

*Time dimension of clustered ODs*

As described in Section 4.1 we cluster ODs based on the first and last decision point in a path. Next, we will have a separate forecasting model for each OD, based on a time series. For this time series to be truthful to the actual data, we cannot just add up all journeys belonging to a clustered OD pair. We need to take into account the time it takes for passengers to arrive at that point.

Based on the time of check in, we will take the earliest arrival time at the clustered origin as an indication of which of the interval the journey should
fall in. For example, when a journey starts at 8:00h and it takes 45 minutes to come to the clustered origin, then the journey will fall into the same forecast interval as journeys starting at 8:45 at the clustered origin. We take the general planned timetable, because we want to predict flows as unaffected yet by the disruption. It is possible to distinguish between arriving passengers and starting passengers at a cluster point.

*Passenger arrival pattern*

We forecast passengers’ arrival in 30 minutes interval, where the start of the interval can be any minute. Aggregating data makes the decision of the start time of the forecast less important. In practice a minute difference in arrival can make the difference between catching a train and waiting for the next. This especially holds true for connecting trains. Moreover, during peak hours more passengers travel. Also discounts have a significant influence on the number of passengers traveling. By aggregating data over a longer period of time, our demand forecasts present a more general estimate of the number of passengers to expect in the coming half an hour. Because the disruption management needs significant time to plan and implement a new timetable and rolling stock schedule, it makes sense to aggregate passenger flows over time. By making the start time flexible, we do capture as much of the current trend in time as possible.

Notice that based on the clustering of ODs and the earliest arrival time, information on the arrival pattern can be deduced. Therefore it is easy to extend the methodology to include a different way of aggregating passengers over time. However, in this paper we do not focus on this specific aspect.

4.3 The influence of different passenger groups

Smart card data stores information on journeys linked to a specific card, which allows for analysis of journey pattern per card and per product type. We need to decide which information needs to be preserved in the forecast, either because it will benefit the accuracy of forecasts or because this information has additional value for the forecasts.

Forecasts distinguishing different types of passengers provide the opportunity to, in follow up research, link reactions to a disruption to a specific passenger type. One can imagine that the reaction of a frequent traveler differs from the occasional leisure traveler, or that the access of student-subscriptions to additional information sources through apps and internet might be better than that of the discount-pensioner card. Hence, distinguishing different passenger types adds value to the forecasts.

Secondly, forecast models might profit from the journey pattern of a card or product type. Possibly, separate forecast models could benefit from this information that would get lost in an aggregation of all journeys.
By deducing travel patterns from the smart card data and investigating the existence of travel patterns per product type, we find an indication of which information may be important for forecasting. We will use these groups in our forecasting procedure in Section 4.4.

We consider the following characteristics for clustering passengers:

- product type (e.g. discount, subscription, regular ticket,...)
- journey class (Dutch trains have two classes. First class is more spacious and more expensive than second class)
- frequency of travel per card
- time of travel per card
- variation in OD per card

For the first two items we can simply analyze whether journeys differ based on product type or journey class over time. To deduce characteristics per card we use data mining.

Data mining travel patterns per card

Deducing patterns in frequency of travel, time of travel and variation in OD, in other words, travel behavior types of passengers, requires data mining. Different journey patterns may be better forecasted by separate forecast models or models profiting from the information of these existing groups.

We use K-means clustering, see Johnson and Wichern (2002), to find the groups of travelers. K-means clustering is an iterative method that minimizes within variance of groups and maximizes variance between groups by changing the location of the centroids of the clusters. Every item in the clustering is assigned to the closest centroid. We use Euclidean distances to measure distance between items. In this clustering, it is very important that all the characteristics of the items (the travelers) are scaled approximately the same. We choose to scale every characteristic to a percentage: either as percentage of the maximum number of journeys of that card, or as a percentage of the maximum over all cards. We use one month of data for which we calculate for every card:
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr of Journeys</td>
<td>The total number of journeys as a percentage of the maximum number of journeys over all cards</td>
</tr>
<tr>
<td>Weekend</td>
<td>Percentage of journeys that were made in a weekend day (Saturday or Sunday)</td>
</tr>
<tr>
<td>Workday</td>
<td>Percentage of journeys that were made on a work day (Monday, Tuesday, Wednesday, Thursday and Friday)</td>
</tr>
<tr>
<td>Different ODs</td>
<td>Number of different journeys in terms of OD combination, as percentage of the maximum number of different journeys over all cards</td>
</tr>
<tr>
<td>Max Time slot 1</td>
<td>Time slot with the maximum number of journeys, as percentage of the full number of journeys</td>
</tr>
<tr>
<td>Max Time slot 2</td>
<td>Time slot with the second-maximum number of journeys, as percentage of the total number or journeys</td>
</tr>
</tbody>
</table>

The above mentioned Time slot refers to specific times of day. According to current NS products, there are two peaks assigned in time: the morning peak from 6h30 until 9h00 and the afternoon peak from 16h00 until 18h30. Some products will allow discounted travel outside these peak hours, resulting in a slight increase in demand before and after the official peak hours. Therefore in terms of time slots, we have decided to define:

<table>
<thead>
<tr>
<th>Time slot</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Morning Peak</td>
<td>5h30 - 6h30</td>
</tr>
<tr>
<td>Morning Peak</td>
<td>6h30 - 9h00</td>
</tr>
<tr>
<td>After Morning Peak</td>
<td>9h00- 10h30</td>
</tr>
<tr>
<td>Afternoon low</td>
<td>10h30- 15h00</td>
</tr>
<tr>
<td>Before Evening Peak</td>
<td>15h00 - 16h00</td>
</tr>
<tr>
<td>Evening Peak</td>
<td>16h00 - 18h30</td>
</tr>
<tr>
<td>After Evening Peak</td>
<td>18h30- 20h00</td>
</tr>
<tr>
<td>Evening/Morning low</td>
<td>20h00 - 5h30</td>
</tr>
</tbody>
</table>

4.4 Prediction Modeling

From Section 4.1 we have the clustered OD groups relevant for the disruption. The procedure from Section 4.3 results in the important passenger groups to be considered. Now we need a model to forecast these OD flows for 30 minutes, 1 hour and 2 hours for any trajectory for any time of any day, based on smart card data up until the day before the disruption.

Special events, holidays, and also disruptions can have a strong effect on the number of passengers traveling. However, we do not want those very temporary events to have a long term effect on our forecasts. Although predicting

\footnote{Different disruptions may have different effects on passenger flows. Hence we do not want a different disruption in the past to influence the demand forecasts of the current disruption. Furthermore note that though aimed at disruption management, our forecasts do not include the reaction of passengers to a disruption.}
these events and their effect is outside the scope of this research, we will focus on the sensitivity of our model to outliers. Consequently we are interested in models that do capture (small) trends, but are not too sensitive to outliers.

We choose to use an autoregressive integrated moving average (ARIMA) model for regression see e.g. Heij et al (2004) and Franses (2004), a widely accepted econometric forecasting method. We use a robust estimation method, comparing results for a cleaning procedure for the data with a robust model estimation using least absolute errors instead of least squared errors.

Furthermore, we investigate whether the groups indicated in Section 4.3 improve forecasts by either estimating separate models for them or by including them as variables in our estimation model.

Finally, we may also estimate the number of passengers per group by using an aggregate prediction model and afterwards splitting this up in a prediction per group. We use a Market Share model (Fok et al, 2001) for splitting the aggregate prediction into multiple categories. This model assures that all estimated fractions are between zero and one and all estimated fractions together add up to unity.

The variables considered for modeling are time, day of the week, autoregressive components and possibly specific weather conditions. All forecasts are based on a real life smart card data set of NS.

5 Results

Our aim is to come to an online forecasting model to predict passenger flows affected by a specific disruption. As this is still work in progress, we present some preliminary results on the data mining process of smart card data. Results of the prediction modeling will be available in the near future.

5.1 Data mining

Using data mining we investigate the importance of the geographical dimension, the product type information and the time dimension stored in smart card data.

5.1.1 The geographical dimension

In Section 4.1 we mentioned that the journeys are not equally distributed over all ODs. Figure 3 shows the cumulative distribution of the number of journeys per OD. The graph indicates that 80 percent of journeys is represented by 15 percent of ODs, and that 5 percent of ODs refer to 60 percent of all journeys. Clearly there is an uneven distribution between ODs and journeys. Still, the ODs with a small volume account for at least 20 percent of all journeys. Hence we cannot just ignore journeys with a low volume. Therefore our approach to
cluster as many ODs as we can without losing important geographical information for disruption management is crucial to come to accurate predictions.

![Graph](image)

Fig. 3: This figure shows that a very limited number of ODs are responsible for the large majority of journeys. 5 percent of ODs contains 60 percent of journeys, 15 percent of ODs contains 80 percent of journeys.

5.1.2 The product type and time dimension

One of the journey characteristics is product type. The product type could also hold information on the journey pattern and therefore improve forecasts when taken into account. Also, forecasts including product type could provide valuable for disruption management, e.g. to enable a focus on the passengers that a public transport operator considers most valuable.

Figure 4 shows the distribution of journeys per product type over time for two different weekdays. The width of the line shows the variation in a month of data. We can see that the pattern per weekday and time unit is highly regular. However, the distribution of journeys over product type is highly dependent on both the day of the week and the time of the day. Also, the total number of passengers traveling (equal to the addition of the lines) is dependent on these two dimensions. Therefore we can conclude that for short term forecasts of passenger flows, the weekday, time of the day and product type can be important predictors for these forecasts.

5.2 Clustering passenger groups based on journey pattern

Smart card data contains information of journeys per card. Possibly, journeys can be better forecasted based on the specific journey pattern of a passenger.
(a) Week day A  
(b) Week day B

Fig. 4: Passengers in the system per product type. We see that different products (lines) have different patterns, and the distribution of the product types over passengers in the system depends on both time and weekday. Graphs do not have equal vertical axis.

Moreover, it can prove valuable to know something about the type of passengers involved in a disruption to model their specific reaction to a disruption. Therefore in this section we use statistical clustering to investigate whether there are passengers with different travel patterns contained in the smart card data, as we described in Section 4.3. We use one month of smart card data for our analysis.

5.2.1 Journeys per card

Fig. 5: This graph gives an overview of the relation between cards (passengers) and journeys. The blue line represents the percentage of journeys (on the y axis) accounted for by the percentage of cards (on the x axis). The red line shows the minimum number of journeys (on the y axis) made by the percentage of cards presented on the x axis. The graph shows that most cards have few journeys while few cards account for most of the journeys.
Figure 5 shows that a small portion of the cards is responsible for the majority of journeys. The red line represents the journeys per card, where the minimum number of journeys per card is decreasing from left to right. The blue line represents the percentage of journeys on the y-axis made by the percentage of card on the x-axis, where percentage of journeys is dependent on the cumulative sum of the cards ordered by decreasing number of journeys. The graph shows that most cards do not have a sufficient number of journeys to analyze travel patterns - almost 70 percent of the cards have made no more than two journeys. However, 60 percent of the journeys are made by 20 percent of the cards. Note that due to the current adoption phase, most regular travelers are still lacking from the data. Hence we expect that in the future an even larger part of journeys will be resulting from passengers that travel regularly.

Travel patterns

We use the statistical package R to cluster the journeys per card based on the frequency of travel, the day of travel, the time during the day of travel and the variation in ODs of journeys. We use k-means clustering. By testing several groups and evaluating the valid number of clusters based on a plot of the within groups sum of squares versus the number of clusters we came to conclude that 4 clusters were present in the smart card data. For more information on k-means clustering see Johnson and Wichern (2002).

The statistics of the clusters are presented in Table 1, and each of the clusters is shortly described in Table 2. The graph in Figure 6 of the clusters, presenting the location of the members of the clusters in respect to the three largest principal components of the data, shows that these four groups can be clearly distinguished. Hence we conclude that there are 4 types of travel patterns among regular travelers at NS.

<table>
<thead>
<tr>
<th>Group</th>
<th>Nr of Journeys</th>
<th>Weekend</th>
<th>Workday</th>
<th>Different ODs</th>
<th>Max Timeslot 1</th>
<th>Max Timeslot 2</th>
<th>Perc of cards in group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.33</td>
<td>0.1</td>
<td>0.90</td>
<td>0.22</td>
<td>0.22</td>
<td>0.34</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>0.42</td>
<td>0.12</td>
<td>0.88</td>
<td>0.44</td>
<td>0.26</td>
<td>0.36</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>0.36</td>
<td>0.08</td>
<td>0.92</td>
<td>0.22</td>
<td>0.36</td>
<td>0.46</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>0.32</td>
<td>0.30</td>
<td>0.70</td>
<td>0.26</td>
<td>0.26</td>
<td>0.37</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 1: Results of K-means clustering: 4 groups are distinguished. Figure 6 shows the three clusters. Table 2 describes the different features of the clusters.

We will use these patterns to investigate whether predictions of passenger flows dependent on passenger type perform better than aggregated predictions over all passengers. We will compare individual models per passenger type with an aggregated prediction combined with a market share model, as we described in Section 4.4. This is however still work in progress.
Fig. 6: Clustering results of cards with a minimum of 20 journeys within a month. Each cluster is presented by a different color. Axes are the three most important components of a factor analysis of the properties of these cards in terms of travel patterns.

6 Conclusion

Rescheduling of rolling stock and the timetable while optimizing passenger service level requires information on real time passenger flows. Since the recent introduction of smart cards, detailed data on passenger flows is available with
Table 2: Results of K-means clustering: 4 groups are distinguished.

<table>
<thead>
<tr>
<th>Group</th>
<th>Color</th>
<th>Description of type of traveler</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>blue</td>
<td>Passengers travel mostly on workdays, time slot seems not of main importance</td>
</tr>
<tr>
<td>2</td>
<td>red</td>
<td>Frequent traveler that also travels in the weekend for whom time slot is important but has most diversity in the choice of OD.</td>
</tr>
<tr>
<td>3</td>
<td>green</td>
<td>Frequent traveler that has a fixed pattern: travels mostly weekdays and in a preferred time slot</td>
</tr>
<tr>
<td>4</td>
<td>black</td>
<td>Of the groups, these passengers travel least frequent, but travel often in the weekend</td>
</tr>
</tbody>
</table>

just one day delay. This data contains information on the card type, origin, destination, start and end time of each journey stored per card id.

We proposed an algorithm that dependent on the disruption transforms smart card data into data suitable for forecasting. This data forms the basis of an online prediction model forecasting real-time passenger flows. Preliminary results showed that:

- Total demand distribution is dependent on weekday and time of the day
- Ticket types show different patterns, both over time and over geographical location
- Passengers have different journey patterns

Further work will focus on estimating forecasting models based on the transformed data set. Furthermore, we aim to implement our algorithm to test our online forecasting model based on a real life case study.

In our current forecasting model, we do not include the reaction of passengers to the disruption. Future research will focus on including passenger behavior into the disruption management process and on using information to influence the reaction of passengers and optimize passenger service level.

Acknowledgements This research was enabled through a grant of The Netherlands Organization for Scientific Research (NWO). Thanks to the collaboration with Netherlands Railways (NS) we were able to use real life data for our analysis. Finally, we would like to thank Dennis Fok, Christiaan Heij and Richard Paap for their ideas and comments on the prediction modeling.

References