Passenger Flows in Disruption Management:
A Dynamic Forecast Model

TRAIL Research School, Delft, October 2012

Authors
Evelien van der Hurk, Gábor Maróti, Leo Kroon, Paul Bouman, Peter Vervest
Rotterdam School of Management, Department of Decision and Information Sciences, Erasmus University Rotterdam, The Netherlands

© 2012 by Evelien van der Hurk, Gábor Maróti, Leo Kroon, Paul Bouman, Peter Vervest
## Contents

Abstract

1 Introduction ................................................................ 1

2 Problem description ....................................................... 1
2.1 Information and Passengers................................................. 2
2.2 The Passenger Guidance Problem .......................................... 2
2.3 Dutch Smart Card Data ..................................................... 3

3 Literature Review .......................................................... 4
3.1 Forecasting OD flows and demand for transport……………… 4
3.2 Research on Smart card data................................................ 5
3.3 Disruption Management .................................................... 6

4 Methodology ................................................................ 6
4.1 Clustering of relevant passenger flows ..................................... 7
4.2 Time.......................................................................... 11
4.3 Prediction Modeling......................................................... 12

5 Results ....................................................................... 12
5.1 Introduction to the data ..................................................... 12
5.2 Forecasting Model and Results ............................................. 13

6 Conclusion .................................................................. 18

References................................................................................ 19
Abstract

Reliability of public transportation systems is important as many people depend on it. Disruption management studies how to improve service level in case of a disruption. Traditionally it focusses most on the resources such as vehicles and crews. We argue that guiding passengers leads to a higher service level, and show that new sources of data, like smart card data, provide a good basis for reliable forecasts of passenger flows. The latter is a necessary condition for including passengers in disruption management.

During a disruption, passengers may end up on non optimal routes due to lack of information. Moreover, vehicles can get due to lack of information overcrowded due to an imbalance between demand and capacity. Both lead to delay for passengers. Accurate and timely route information enables passengers to chose better routes, reducing delay for themselves as well as as the system.

This research is about guiding passengers and balancing demand and capacity through information. Passengers are assumed to freely chose their own route in the system. Hence the influence of information is strictly limited to a set of options that each traveler individually considers acceptable.

The paper proposes a type of multi commodity flow model to solve the information problem. Here, the network is defined by available capacity, the commodities represent the time dependent Origin Destination flows and the behavioral aspects are included as restrictions in the model. As information on passengers is essential for solving this model, the paper focusses on forecasting passenger flows from new data sources as smart card data by a dynamic forecasting model, defined in this paper.

First results indicate that the application of this newly proposed algorithm to a big data set like smart card data generates reliable estimates for the flows in the network, which can be used as input for the optimization. Results are based on real life smart card data. Focussing on the influence of passengers on system service level is a new approach to disruption management, that can compliment current research focussed on capacity management. Combined approaches will lead to more reliable transportation systems. Though the focus in this paper is on public transport, we strongly believe that similar models can be applied in traffic as well.

Keywords

Passengers, Disruption Management, Forecasts, Econometrics, Optimization
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, accidents and technical problems are just a few examples of the causes of disruptions. Disruption Management aims to improve service in such cases.

When a disruption occurs, complex rescheduling problems need to be solved quickly. Though some research on adjusting the logistic schedule to the new situation exists, there has been not much attention for influencing passenger’s path choice. However, passengers’ route choice determines the delay of both themselves and others, as balance between capacity and demand is an important factor in the incurred delay. Accurate information enables passengers to make better route choice decisions, and guidance of passengers may reduce or prevent overcrowding of trains.

How to distribute passengers over the network by providing them with information is a complex problem. This paper proposes a formulation of this Passenger Guidance Problem (PGP) as a type of multi commodity flow model. The network is described by the timetable, the capacity restrictions by the capacity of the vehicles. The commodities represent the passengers. Restriction on the paths of the commodities represent behavioral restrictions.

Detailed short term forecasts of passenger flows are needed as input for the (PGP). 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 and time of departure.

Until recently, detailed information on passengers’ journeys was not available. This has recently changed due to the introduction of smart card data. This paper presents 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.

First results on real life smart card data of Netherlands Railways (NS) are presented in this paper. Results indicate that based on past data accurate predictions of passenger flows can be obtained from smart card data, as needed for the PGP.

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: The timetable needs to be adjusted, rolling stock needs to be rescheduled and new crew duties need to be generated. All of this needs to happen quickly, while often not knowing how long the disruption will last.
At the same time these logistic problems need to be solved, 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.

The research question we address in this paper is:

*How to optimize passenger service by rerouting passengers in case of a disruption?*

### 2.1 Information and Passengers

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.

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. In practice, a two yearly survey is used to estimate OD matrices in general. For operational planning an advanced econometric model is in development to estimate passengers per train. During a disruption, passenger flows are based on expert opinion.

Only since the recent introduction of smart card ticketing, the time, ticket type, origin and destination of each journey is registered. In the Netherlands, smart card data is available to the Public Transport Operator with one day delay. Hence it is a great source for the short term forecasts needed to balance capacity and passengers in case of a disruption.

Furthermore, technological advancement makes it now possible to communicate with passengers at any time, anywhere. Thanks to internet and cell phones there are plenty of opportunities to provide passengers with information. With timely accurate information delay of passengers can be reduced because it enables an early change of route which can lead to less delay. Moreover, route advice can be used to spread passengers over the network and balance capacity and demand.

### 2.2 The Passenger Guidance Problem

We address the problem of balancing capacity and demand from an operator perspective, where given the newly available sources of information a strategy for providing route advice to passengers needs to be generated. As the problem is about guiding passengers we refer to this as the Passenger Guidance Problem(*PGP*).

We can see the *PGP* as a type of multi commodity flow problem, where:

- the commodities are the passengers
• capacity restrictions are defined by the rolling stock schedule
• the network is the timetable
• we minimize overall delay of passengers
• behavior of passengers can be included as restrictions on the routes of specific flows

A similar approach was taken in Nielsen (2011) where an assignment of passengers to trains was used to optimize capacity. However, at that time no forecasts of passenger flows were available. This paper is about forecasting the information on passengers for disruption management. We focus on how to include passengers in the model for balancing capacity and passengers.

Including single passengers would lead to many variables and probably would make an accurate forecast of passengers very hard. This is because although information from smart card data is available per card - and hence per passenger - many passengers travel infrequently. Moreover, many passengers do not have an easily predictable pattern into their journeys.

The PGP is about distributing passengers over the network in a way that increases service and reduces delay of passengers. The relevant routing options depend on the Origin and Destination of a passenger (the OD). Moreover the time of traveling is important in a system that works with a timetable like the Dutch system, with train services between one to 6 times an hour. All route choices based on these three aspects are similar for all passengers. Therefore passengers are modeled by OD variables, representing Origin, Destination and start time of the journey.

Our Dynamic Forecast Model is focussed on generating short term online forecasts of these OD. Therefore we use two insights into the problem. One is that for providing information, focus should be on passengers affected by the disruption. Second is that different OD’s may have similar choices available to them. In fact, the important points (stations) in the network are those where passengers of an OD need to make their first routing decision due to the disruption.

Clustering OD’s based on these decision points has many advantages. It provides insight into the way the disruption affects the network. It gives information on when and where passengers need to have information on route choice. Aggregating these OD’s can have advantages in forecasting. Finally, it is a good preprocessing step to reduce the number of commodities needed for the multi commodity flow model.

2.3 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 passengers, 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 & 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.

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 & 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 & 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.

\(^1\)AURORA is the name of a series of projects focused on long term predictions of the number of passengers per train.
**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 focusses 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

The Dynamic Forecasting Model has two phases. First, we convert smart card data into several time series, second we use these time series for prediction modeling. In the first phase journeys similar in route choice are aggregated into a time series, by aggregating journeys (Section 4.1 and discretizing time (Section 4.2) leading to the \( OD_t \) values in
Section 4.3 describes the econometric model used to forecast the number of journeys.

### 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 loose 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 1a 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...
(a) Find relevant OD flows through shortest path computations

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

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

**Figure 1:** 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.

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 1b.

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 1c. The first is MKJHG and the detour path is MKJIAHG. 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, 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 $MKJIABG$ 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 ($D$ to $\tilde{D}$). 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 $\tilde{O}\tilde{D}$ combination, we only need to find the decision points and not the path itself.

We define:

- $O'D'$: a specific OD pair from the set of all ODs
- $\tilde{O}\tilde{D}$: a pair of two stations that possibly represents multiple OD flows
- $d(O'D', N', X)$: the journey time from $O'$ to $D'$ in network $N'$, either through path $X$ or through a shortest path if no $X$ is provided
- $p(O'D', N', X)$: the order of stations in the path from $O'$ to $D'$ in network $N'$, either through path $X$ or shortest path if no $X$ 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 $O'D'$ from the set of ODs. Set $N$ and $N_k$ to initial values.
2. Set path $A$ as $p(O',D', N)$ and path $B$ as the shortest path in the network going through any part of $k$ $p(O'D', N)$
   - If $d(O',D', N, B) > d(O',D', N, A) + \alpha$, go to step 1
   - Else go to step 3
3. Set path $C$ as $p(O',D', N_k)$. Set the Minimal Detour Time $S$ equal to $d(O',D', N_k)$.
   - If $S > d(O',D', N, A) + \beta$, then add $O'D'$ to the set of journeys for which no reasonable detours exist. Go to step 1.
   - Else go to step 4
4. Compare $p(O',D', N, B)$ and $p(O',D', N_k, C)$. Set $\tilde{O}$ equal to the first point where the paths start to differ and $\tilde{D}$ to the latest point at which the paths are first joined.
   - Else if: $O'D' = \tilde{O}\tilde{D}$, add $O'D'$ directly to the dictionary. Go to step 3.
   - Else go to step 5
5. Delete $\tilde{O}$ from the network $N_k$. 
6. Set path $E$ as $p(O'D', N_k)$.
   - If $d(O'D', N_k, E) \leq S + \gamma$ path $E$ is a reasonable path, set path $C = path E$, repeat step 4
   - Else go to step 7

7. Fix $O' = \tilde{O}$. Restore $N_k$ to initial values, then Delete $\tilde{D}$ from the network $N_k$.

8. Set path $F$ as $p(O'D', N_k)$
   - If $d(O'D', N_k, F) \leq S + \gamma$ this new path is a reasonable path, set path $C = path F$, 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 - leading to an almost continuous dimension of time. Time series analysis requires discrete time blocks. Therefore passenger flows need to be aggregated over time.

Time is important in our model. Passenger flows are strongly dependent on the time of day, as peaks results from passengers commuting. Moreover, in a timetable based system, the opportunity to board a train or to transfer from one service to another is of string influence of the travel time. For example in the Dutch system, this difference may be between 10 to 60 minutes, and will most commonly be around 15 minutes. Finally time is important for balancing capacity and passengers - as estimates of the number of passengers per train are required for this.

Therefore, we make aggregation of time dependent on the train service a passenger can connect to. Descretizing time on the clustered origin nodes enables a focus on when and where balance between capacity and passengers becomes important, as this is the first point where passengers may change their route.

Two situations are distinguished. One where passengers arrive at the clustered origin as the start of their journey: they are aggregated based on the originally planned trains they would take in case there was no disruption. A second where passengers arrive by train to the clustered origin: they are clustered on the arrival time of the train at this clustered origin, as from this time, they can change there route.

The latter needs assignment of passengers to a specific route. As 70 to 80 percent of journeys has a direct connection, and he clustered origin forms the first point where a actual route choice needs to be made, we expect that assigning passengers to the trains on their shortest route will have sufficient accuracy for our models.
4.3 Prediction Modeling

The clustering algorithm in 4.1 and the time aggregation procedure in Section 4.2 together transform a set of smart card data to a set of time series. These time series correspond to past $OD_t$ flows in the model for $PGP$, presented in Section 2. With timeseries modeling, these time series containing information of the past are used to generate forecasts of future passenger flows.

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.

The general model is:

$$y_t = c + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \beta_{p+1} \varepsilon_{t-1} + \beta_{p+q} \varepsilon_{t-q}$$  \hspace{1cm} (1)

$y_t$ The time series, which can also be integrated.
$\varepsilon_t$ The error at time $t$
$c$ A constant
$\beta$ Coefficients

An analysis of autocorrelation decides the integration factor used. Analysis of a subset of OD-flows will decide the ARIMA level to be used (the number of past terms of the series, as well as the number of error terms).

5 Results

Results are based on a 4 months smart card data set from Netherlands Railways, the largest passenger railway operator in the Netherlands. As introduction of the system is currently not finished, data contains about 50 percent if all journeys. As full introduction of the smart card system is expected in the near future, we will in this research not go into scaling, but focus on the data set as is.

Results focus on the quality of forecasts of passenger flows based on smart card data. The ambition us to come to an integrated online model for guiding passengers in a disruption through information, balancing capacity and demand. As this is still work in progress, we present some preliminary results now - we are currently working on a full case study, and expect to be able to present these results in the very near future.

5.1 Introduction to the data

This Section provides insight into the smart card data before we go into the forecasting results in Section 5.

Figure 2 shows the cumulative distribution of the number of journeys per OD. Clearly the volume of journeys is unequally distributed over all OD pairs as about 80 percent of journeys is represented by 15 percent of ODs. Still, the ODs with a small volume
account for at least 20 percent of all journeys. It is important to take them into account this twenty percent, as they are unequally distributed over the network. Omitting them could therefore have serious impact on forecast accuracy for some parts of the network.

![Number of journeys over ODs](image)

**Figure 2:** 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.

In Section 2.4.2 we mentioned the importance of time as passenger flows are time dependent. This pattern is shown in 3. The graph displays the number of people traveling over time, the axis has a length of 24 hours. Each differently colored line is a weekday. The amplitude of the line refers to the variation in the number of people traveling for that specific time of day. Results are for one month of smart card data. Figure 3 shows that not only is the volume of passengers traveling dependent on time, it also depends on the day of the week. Given the time of day and weekday, this volume is rather stable - hence taking into account these variables can greatly improve forecasts in comparison to a yearly average, as is current practice.

### 5.2 Forecasting Model and Results

The number of passengers traveling per day on aggregate as very regular, as shown in Figure 4 and is supported by the results in Section 5.1. The outliers in figure 4 can all be explained by special events like Christian festivities and school holidays.

This Section presents results for modeling and forecasting individual ODs per day by a time series model as described in Section 4.3.

In the ARIMA model the integration factor (I), the autoregressive component (usually referred to as $p$) and the moving average component (referred to as $q$) need to be decided. It is important that the series is stationary. The autocorrelation series of a time series is usually a good indicator for what a good integration level would be to obtain a stationary series.

It turns out that all time series in our sample of 8 have a similar pattern in autocorrelation. In figure 5 this pattern is shown by the black line with the high amplitude: a very
Figure 3: Number of passengers over time per day. Different colored lines are weekdays, amplitude is variation in number of passengers that travel within a specific month.

Figure 4: Total number of journeys per workday over 6 months - weekends show similar pattern though they have less journeys.
Figure 5: Autocorrelation of the time series (black) and autocorrelation with lag 7 (blue) for one OD. The original time series has a high amplitude with period 7, indicating a high correlation between weekdays. When differencing by 7, the autocorrelation is strongly reduced and dies out.

A strong autocorrelation exists between values 7 days apart. Even further than that, in lags of 7, autocorrelation remains high. This can be explained from the high regularity between days of the week as shown in Figure 3.

We differenced the time series by 7. The autocorrelation of this lag-7 time series is represented by the blue line in Figure 5 which has far less amplitude than the non-differenced time series. Also the series appears to be stationary after this integration, because the 7-lag differences colored by weekday in Figure 6 are similar over the days and show no clear trends. Remember that in the original series the number of journeys strongly depended on the weekday - leading to heteroskedasticity. By differencing this problem has gone: meaning that although the number of trips depends on the weekday, the variation on the number of trips from week to week is similar for all days.

Because of the autocorrelation series, we chose to estimate an AR(1) model with $i = 7$. A short exploration to include up to 5 autocorrelation terms and up to 5 error terms did not indicate an improvement of results in comparison to the AR(1) model - hence we will present the results of the AR(1) model only.

The AR(1) model captures some of the variation in the time series, but has difficulty with large variations. In Figure 7 the red line representing the forecasts of the model follows the general pattern of the black line (the original forecasts), but especially at the end of the series where variation is higher the error increases. Possibly by finding what causes the large variation in the original series, the forecasting model can be improved: either by including variables that explain the variation or by adjusting the time series, e.g. exclude outliers.

For 8 different OD flows we have estimated the model. We chose the ODs such that they would be aggregated in case of at least one specific disruption. We have one sample with five fairly small flows, and one sample with one large flow and two smaller flows.
Figure 6: By taking lag-7 of the time series, the series appears stationary: the data vary around zero without a clear trend or clear differences between weekdays.

Figure 7: The differenced time series (black) and the forecasts (red line) for one OD.
Each OD flow has a separate model, furthermore, two aggregated models are estimated. Constants for the model were not significant, which is an indication that there is no unit root in the series which is important for the reliability of the regression results.

In Table 8 per OD the coefficient of the $y_{t-7} - y_{t-1}$ term is represented together with standard deviation - all coefficients are significant, have the same sign, and most of them are around the same value. Exception is ‘station one’ of series A, but this OD has a very small number of trips - only 9 on average. The aggregated models have similar coefficients. This supports the idea that flows can be aggregated for forecasting without losing important information of individual flows, as there appears to be a rather general model.

The variation explained by the model for the differenced series is around 15 percent. When converting the forecasts to he total number of trips, the variation explained of the total number of trips is far higher. When focusing on a semi-R-squared, an R-squared calculated for the original series and the forecasts converted to total number of journeys, especially for the aggregated series this is quiet good 70 and 83 percent. This indicates that last weeks number of journeys are already a very good indicator of the number of journeys this week, and we can improve this estimate slightly by using a timeseries model.

By aggregating the timeseries, the reliability of the forecasts stays rather stable. The standard deviation of the coefficients is fairly similar over all models - individual and aggregate - which indicate no difference in reliability of forecasts.

When we consider the journeys per OD per day as random variables drawn from a stochastic distribution, the aggregated model would represent the sum of these distributions. The forecast would be the expectation of that distribution, and the reliability corresponds with the resulting standard deviation. The standard deviation of the summed distribution corresponds to the sum of all standard deviations of the individual distributions and the correlations between the distributions. When the series are independent, the standard deviation of the summation is the same as the sum of all standard deviations. When there is negative correlation, the standard deviation reduces, and vise versa for positive correlation.

As in current results the standard deviation of forecasts for the sum of the series is equal to that in the aggregated model, this indicates that series are independent. For independent series, aggregation will not lead to higher reliability. Still, correlation between flows could be expected as well. Given that the number of journeys in the system is rather stable, one OD attracting more trips could mean that another will have less trips, resulting in negative correlation. Positive correlation could also be plausible reasoning from a purpose of travel perspective: when one destination is attractive, for instance because of a special event, flows with that same destination which are aggregated are likely more positively correlated.

Clustered ODs are similar in path choices available to them. Therefore there is no advantage in including them separately in the PGP. For balancing capacity and demand it is about the total flow that is assigned to a specific train. By estimating an aggregated model, we do not need to worry about possible correlation between series that affects the reliability of the forecasts. Hence estimating an aggregated model, even if it does not increase reliability, has an advantage over estimating individual models and aggregating them later on.
Figure 8: Comparison of individual and aggregated forecast performance: A) five small OD flows; B) one large, two small OD flows. Given are the coefficient of the first autocorrelation term of the differenced series, the standard deviation, the R-squared, the semi R-squared for the total number of trips, and the average of the differenced time series per OD. Models are fairly similar over ODs. Aggregation does not lead to an improvement of reliability, indicating independent series.

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 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.

This paper proposed a model for balancing capacity and passengers by providing them timely with accurate information: the Passenger Guidance Problem. The definition is based on a type of multi commodity flow model, where we model passenger behavior by restricting the routes specific flows can be assigned to.

Accurate information on passenger flows is essential for solving this problem. Therefore this paper defines a Dynamic Forecasting Model that, based on a specific disruption, generates forecasts of passenger flows from smart card data. Our results show that there is a strong correlation between passengers per day, and that flows are time dependent. Furthermore, first results indicate that smart card data can deliver accurate short term forecasts of passenger flows by using time series modeling.

By using our Dynamic Forecasting Model we are able to obtain insight into a disruption and reduce the number of forecasts needed. This is a great improvement from current practice, where information on passenger flows results from yearly surveys and expert knowledge. The weekly and time variation of passenger flows is not captured in these traditional information sources, while the analysis and models in this paper show that they are important in forecasting passenger flows.
We are currently working on a case study using the here presented models, which we expect to present in the very near future.

References


