Predicting travel volumes for long-distance coach services through big data analytics - a case study on German public viewing events during the UEFA EURO 2016

Corinna Fohrholz
University of Potsdam
cfohrholz@wi.uni-potsdam.de

Christian Glaschke
University of Potsdam
emailaddress@xxx.xxx

Abstract
Capacity planning is a key challenge for long-distance coach vendors due to the competitive market pressure. Travel-intensive events have the potential to increase the capacities for existing routes or to offer new connections. Actually, planning activities of vendors are mainly based on past-data. To improve the quality of predictions about future capacities, information from different data sources e.g. social media, geo-based information or websites has to be considered. Technologies to collect structured and unstructured information from different data sources are available. This paper presents an approach to use these technologies for capacity predictions and pricing of coach vendors in terms of travel-intensive events. To verify the approach, a case study, focused on the UEFA EURO 2016, is described. The case study is necessary to demonstrate the value of the research and to give a deeper understanding about data sources and the discussion culture of travelers in social networks.

1. Introduction and Motivation
The market for long-distance coach travel in Germany has been highly competitive, since deregulation prompted a number of providers entering the market [1]. Moreover, the image of coach tours has changed owing to these new providers, whereby increasingly more people in Germany can imagine travelling by coach. Therefore, companies need to offer new formats to further expand their market share. Travel-intensive events such as festivals, city festivals, fairs and demonstrations offer potential in this respect, with statistics highlighting the popularity of such events among the population: indeed, almost 25% of people in Germany are interested in major events such as rock concerts, festivals or public viewings [2]. The following example shows, that even short-term occurring events have an impact on travelling activities. On 9th July 2014, the German soccer team won the semi-final against Brazil and made their way into the final four days later. One hundred thousand people watched the final game at public viewing points across the country. Especially Berlin was a destination for soccer fans. Two days later, a fan festival was held to receive the world champions, with fans travelling from all over Germany. Several hundred thousand visitors watched the game and welcomed the team.

Planning the capacity of long-distance bus suppliers is undertaken at present because it is mainly based upon information from the past, as well as forecasts of a possible utilization. The demand for connections may change according to each date in terms of the volume, periods and geographical distribution and willingness to pay. Especially for events where many people come together for a short period, an accurate prediction of the requirements is necessary to avoid derating [3]. False predictions lead to either lost profits or inflated costs. Social network analysis offers an opportunity to obtain more information about travel behaviors and potential customers. In social networks, announced and discussed events contain information about the numbers of participants and target group characteristics [4][5]. In forums and blogs such as Facebook and Twitter, users exchange travel plans, including time and transport options [6]. For individual events such as festivals, there are even dedicated discussion forums.

Therefore, this paper introduces a concept to support demand planning activities and price-preferences from long-distances coach companies through targeted social media analysis. By that, companies are able to perform quick and customer-oriented planning processes, especially for travel-intensive events. For the evaluation of the overall concept, it is necessary to perform case studies, whereby the focus is on major sporting events. Sports
tourism in general plays an important role in Germany [7]. Fans travel with their teams in their spare time and even support them at away games [8]. As the most popular sport in Germany, soccer comes particularly into focus. During major competitions in recent years, public viewing has gained enormous importance. To validate the concept it is planned to conduct a first case study during the UEFA EURO 2016. The aim is to analyze travel behaviors of social-media users during the games through social network analysis. The case-study is going to be complemented through a questioning of public-viewing visitors. Based on the available data, a customer forecast for selected long-distance routes is going to be carried out. Forecast and the actual utilization are compared with each other. The difference shows the potential of the social media analysis. Firstly, this paper discusses the findings of the case study and the further procedure is explained.

2. Current state of research

2.1. Sport and event tourism

There are different approaches to classify events with different indicators. Getz [9] concluded that a definition of events is always subjective, because even a festival can have a significant impact on the tourism sector or the economic income of a small town [10]. As an example, each year in June, more than 80,000 visitors travel to a small town called Mending in Germany to visit the “Rock am Ring” festival [11]. Around 8,000 people live in Mending.

According to Weed [12], events can be classified into three main groups: business events, cultural events and sporting events. They are also classified depending on their size into minor, major, hallmark, mega-events and festivals [13]. A mega event is short-term and has a fixed duration [14]. Hallmark events are major one-time or on regular events with an international status [15] and they are a permanent institution in the community [10]. Regardless of the size, sporting events can be divided into three main groups: regular events held at different places (e.g. FIFA World Cup), events with a strong association to a place (e.g. Wimbledon) or franchises with general spatial stability [16]. There are examples of all such kinds of events around the world. Sporting events are not only relevant for the tourism industry; indeed, studies have shown that tourists are willing to travel long distances to attend a sport event. Additionally, they spend money for accommodation and other activities around the event [16]. Moreover, they also improve cultural offer and the image of a city [17]. If a mega event is hosted by a city or region, there are positive side effects like recreation and urban regeneration projects. For a city or region hosting a large event, handling the traffic is an important consideration [18]. Accordingly, different industries can benefit from such events, although it is not necessarily ensured that a country or region profits from such an event in economic terms. Indeed, much research is available about the economic impact of large sport events [19] [20] and amateur sporting events [21].

For the 2006 World Cup, hosted in Germany, researchers subsequently found out that the impact on German GDP was less than 0,02% [22]. Jobs were only created in the low-paid sector. However, there are some facts that cannot be measured in Euros or an increase in GDP; for instance, the 2006 World Cup was one of the largest media events ever, with more than 26.3 billion people worldwide watching the games on television (Fifa Infoplus [23]). Baasch [22] provides an overview of positive and negative effects in his work. The phenomenon of public viewing (sometimes also called FIFA fanfest) occurred during the FIFA World Cup in 1998 and has been expanded during the following tournaments in 2002, 2006 until today. Public viewing during large sport events has become very popular, especially in European cities [24]. At the 2010 World Cup, 16 million people watched the game between Germany and Uruguay in the form of public viewing [25]. A total of 350,000 fans attended the International FIFA Fan Fest in Berlin for the semi-final match between Germany and Spain (Fifa 2015 [26]). During the World Cup in 2006, hosted by Germany, Preuß et al. [27] conducted an in-depth survey with 9,456 stadium and public viewing visitors, concluding that public viewing offers were not only attended by locals. Indeed, more than 20% of the visitors of public viewing defined themselves as World Cup tourists and 21% were so-called “time switchers” (people who changed their holiday plans because of the games). Especially the latter had no admission tickets for the games themselves.

The Federal Statistical Office Germany states that more than 20% of the German population is interested in attending public viewings [2]. Nearly 25% [2] are interested in large events in general (including rock concerts, festivals or fairs). In the U.S. public viewing is not common; rather, large
sporting events are usually watched in sports bars or in front of a private TV [25]. However, not only big sport events like FIFA World Cups lead to travel activities of fans. Phil Turbutt [28] concluded in his research that English football tourists are driven by a combination of leisure motivation, location, behavioral and socio-demographic traits and the event itself. For the British Open in 1999, more than 70 percent of the visitors were tourists from outside the community [19]. While most of the studies discussed are focused on large sport events, there are also some impacts of studies about smaller events. Daniels and Norman [20] examined seven local sport events. Their results show that sport events are a suitable way to bring visitors to a site that they might not otherwise visit. In five of the seven case studies, the participants indicated that they visited the host country primary to attend the sport event. Soccer is highly valued in many European countries. Each weekend, thousands of fans travel to follow their teams and support them at away games. In the 2013/14 season, 13 million people attended soccer games in the “Bundesliga” [29]. There are also cases, whereby the tourists attend an event as an athlete e.g. the marathons in large cities. Each year in April, one of the largest marathons in Europe, is helded in Paris. In 2015, more than 42% of the runners were foreigners, most of them from the UK [30]. A running fair two days before the marathon also gathered 80.000 visitors [31].

2.2. Long-distance coach services travelling

As stated in the previous chapter, there is a huge market for public transportation around different types of sport events. The long-distance passenger transport system includes several components: high-speed rail, airlines, long-distance coach services and the car. Due to rather short routes in Germany in comparison to the US, the long-distance coach service is in competition with a proper motorway and railway network. According to the World Tourism Organization a long journey comprises a minimum of a 100km journey [32]. Coach traveling can be divided into three categories: long-distance, coach tours and hire coaches [33]. Long-distance coach services are characterized as offering transportation outside the urban areas, usually from city to city [34]. Destination is the main criteria when talking about long-distance coach services, because it determines the local regulations [34]. The term “coach” is defined as a road motor vehicle that is exclusively constructed for the carriage of seated passengers [35]. By comparison, a bus is constructed in a way that seated and standings passengers can be transported [35].

In comparison to other countries like the UK or Sweden, long-distance coach services in Germany were heavily restricted [36]. The deregulation of the long-distance coach market in Germany was implemented in 2013. Prior to which the demand for long-distance coach services was around 3 million customers each year, whereby Deutsche Bahn AG had a monopole on long-distance travelling. In 2013, the demand increased up to 8,2 million [37]. Shortly after the deregulation, 143 routes had been available. In April 2015 this number had increased to 264 routes overall [38]. Privately owned, mostly small and medium-sized start ups have entered and brought innovative ideas and dynamics into the market [1]. There are still some restrictions to protect local public transport in Germany; for instance, a minimum distance of 50 km is necessary between stops of a coach service [39]. Furthermore, it is forbidden to offer coach services if there is a parallel regional rail transport with journey times up to one hour. The growth potential of the market is still considerable, because there are still routes where no regular rail connection exists and the distances are much lower in comparison to the U.S. [1]. Almost 40% of the German population state that they are interested in using long-distances coach services [40]. On the other hand, 35% are not interested in the service. However, the coach market in Germany is still changing. As other examples have shown, after a period of market opening, the share of many small and medium-sized and a few large companies reversed and mergers and acquisitions eliminated new entrants in the market [41]. In comparison to other European countries, the German market is still very small. In the UK, the largest vendor for long-distances coach services serves almost 1.000 destinations [33]. In the European Union, the regulations range from being very liberalized to prohibition [42] [43]. According to a survey, visiting friends and holiday purposes are the most common reasons for long-distance travel [44]. It is difficult to define the reasons for travelling, given that they are related to each individual’s socio-economic and psychological conditions, as well as background characteristics [45]. In Great Britian, coach travel accounts for 6% of all long-distance milages. Long-distances coach services are becoming more popular, although holiday and leisure travel are more price sensitive than business and commuter travelling [46]. In America, coach travel is very popular but it is associated with a negative image, because it is used especially by low-income groups and older people.
In comparison to Europe, the railway network in America is less strongly developed. On the other hand, coaches connect more than 15,000 communities [47]. The prices for long-distance coach travelling in Germany are still very low, averaging almost 14 Euros per 100km [41], in comparison to railway or even plane. Large vendors use this strategy to step into the market, gaining shares especially in the first years after the deregulation. Indeed, this strategy succeeded: two years after the deregulation, the number of vendors decreased. At the beginning of 2015, Flixbus and MeinFernbus - the two largest vendors in Germany - had a market share of nearly 57% [48].

Research about transport preferences of event tourism and sport tourism in particular is rare [3]. Existing research is focused on the form of long-distance travel in General [46], coach tours as a form of holiday or transportation concepts around large events [3]. Mason and Beaumont-Kerridge analyzed the form of transportation during popular open-air music festivals in the UK, concluding that most of the visitors used a car [49]. Some statistics about preferred transportation of sport tourists during the FIFA 2006 World Cup in Germany are available [50]: most people (57%) used their own car, railway (36%) and plane (24%), whereas only 5% arrived by using a coach service. It is interesting that there is a difference between German and foreign guests, whereby twice as many foreign guest used coaches to reach their destination. If the reason for travelling is to attend an event, then the demand is constrained both in time and space, as well as the price. Visitors try to reach their destination with the best combination of travel time and price. The inability to provide additional capacity to meet peaks in demand for very short periods is the core feature of many transport dilemmas and applies equally to private transport by car and to public transport operators [3]. Conversely, the vendors of transportation services have to meet the demand considering their own capacities. Cost coverage is a key criteria before offering additional capacities. Especially in the highly competitive market of long-distance travelling, vendors need strong tools to predict their capacities.

3. Prediction of Travel volumes

3.1. Former Research

A prediction is a statement about the way things will happen in the future [51]. Research about different prediction approaches is available. In terms of travelling, most research is about the prediction of "traffic conditions" and "travel time" [4] [5], whereas other authors deal with the prediction of traffic volume, especially for short distances in urban areas. This paper is focused on the demand for long-distances coach services. In this specific research area, the work of Yang [6] can be mentioned. It deals with the forecast of visitor numbers for a Chinese holiday destination. For this purpose, the requests are retrieved from search engines to determine the frequency of a search term. The sum of the queries is subsequently used to determine the popularity of the region and thus provide an indication for travel volume. Another application for prediction is that the travel industry is trying to combine various sources of data including social networking data to create personalized vacation packages for their customers [52]. All authors used different data sources and new technologies in order to make a prediction. Such a prediction depends from the event (region and of course reputation), the expected traffic (volume) and the capacity itself. The research gap is determined by the particular application of long-distance coach services, especially for event tourism. The previous work, presented in this paper, does not consider the prediction of travel volume for large or mega events. Because of that, customized offers for long-distance coach services are usually based on data from the past. Short-term demand peaks, such as the entry of the German soccer team in the final of the FIFA World Cup 2014, are a special challenge for vendors. There are just a few days between the announcement and the event (5 days in the case of the 2014 World Cup). To predict the demand as accurately as possible and as soon as possible is a key challenge for vendors. The second challenge is the decision if an additional coach route will lead to more earnings.

3.2. Prediction model

To predict the travel volume for large or mega events, a prediction model, which is shown in Figure 1, has been developed. In general, social networks are a suitable source to estimate the popularity of an event because information about potential customers are available. Most large events have own Facebook pages and Twitter accounts, or members create own groups/blogs, to exchange ideas and meeting arrangements. For the prediction model we suggest the following indicators. (1) The membership in an event related forum is a weak indicator for attending the event [53]. But, a response to an event page or forum does not automatically mean an actual attendance offline [54]. Former research indicates that the public nature of the event, the invitee list and
the visibility of the participants affects the visit of the offline event. The authors of that study argue that it is easy for an individual to click a button to join the event. But in comparison to indicator (1), joining an event „online“ is an indicator that a person is going to attend the event (2). And it is stronger than just the membership in a group (3). The number of event related comments and activities in social media is a strong indicator that a person is interested in the topic and therefore is going to attend the event. In terms of political activities several studies had examined the relationship between online activity and election outcome [55] or in attending protest during the arab spring [56]. It can be assumed that these findings may be transferred to sport or other mega events. In the german speaking Facebook community, more than 50 public groups are available. In the structural equation model, the four indicators are summarized as the „social media factor“ (4).

Figure 1: Structural equation model

The user itself is the second factor in the prediction model. The age is one of the key factors when talking about attending mega events. A study, during the Olympic games in Australia found out that younger people are more interested to attend sport events [57]. On the other hand, expenses for attending sport events increase with the age [58]. In their work Preuss et al. couldn’t find a significant correlation between gender and the attendance of sport tourism [27]. But male visitors spent more money during public viewing events at UEFA WM 2014. Because of missing research in that area, the gender is part of the prediction model until the pre-validation (see chapter 4) is finished. The last indicator for the demographic factor is the social status. Profiles in social networks often contain information which can be used to gain conclusions about the social status. This can be for example job descriptions or the employer is named in the profile. People with lower income, for example students, are more price sensitive than employees. Expenditures for travel, and expenditures for events, rise with an increase in income. Both factors, social media factor (4) and demographic factor (8) influences the actual attendance of the event by a person (9). There are more factors with an impact on the decision, but the ones named above can be determined through social media analysis, which is going to be described in the next chapter.

As mentioned in the previous chapter, 40% of the german population is generally interested in using long-distances coach services [40]. Transferred to the prediction model, this implies that at least 40% of the identified users are potential customers for long-distance coach travelling. But the choice of the transport method is determined by several influence factors. The correlation between age, gender, social status and the ability for long-distance coach travelling has been identified in previous studies [59]. Because of that, the demographic factor influences the intention to use long-distance coach services. Another indicator (12) is the travel time in relation to the price [60]. The statistical federation Germany asked potential customers about their attitude against long-distance coach travelling. Around 35% of all people in Germany, mostly high earners, cannot imagine to use a coach [40]. That has to be considered in the model (11). The last two indicators are the availability of alternative travel method (14) (e.g. own car or train) and of course the general availability of a coach route (13). All these factors are part of the prediction model and has to be collected during web analysis.

3.2. Technical implementation

The main challenge is to process and analyze social media data for planning long-distance couch services. Existing solutions, such as data warehouse or statistical software, are not able to process large and unstructured datasets within a reasonable amount of time [61]. The second main challenge is to collect huge amount of unstructured data through the usage of a web crawler, automatically. Analyzing the data and the identification of all relevant information for the structural equation model, is the third challenge.

The first step is the web crawler implementation. For that, a list of potential websites and login details is necessary. Secondly, the crawler needs information about the website structure. That means, the crawler has to learn the positions of comments, user names and any other information which are needed. To gain information from different social channels, web
several crawler instances are necessary. Each website (e.g. social media channel or network) is represented by an own crawler. For each event, the system needs some basic information like place, date and length. The crawler has to be tailored in terms of starting page and keywords to search for. The crawler starts to search for user by using the website scheme. The web crawler tries to determine all information for determining the social media and demographical factor (e.g. user names, gender and comments). Additionally, users leave digital footprints for identification and localization while using social networks [62]. These information are necessary to built regional clusters in the second step. All information which are collected, are going to be stored in a database (see figure 2). The database stores the data from different all web crawler instances. Any changes are transmitted in real-time.

Each identified user is represented by a user profile in the database. It happens, that users are active on several social media platforms. Because of that, an algorithm for identification and elimination of duplicates is necessary. Existing algorithms for similarity analysis are going to be used.

According to the structural equitation model a user ranking is created. The databases contains information about standard routes of long-distance coach service vendors. These information are going to be compared with geo-information in the user profile. Based on the available geo-information about the user, the systems build regional clusters. The user is assigned to the cluster with the smallest distance to a stop on a coach route. The result are regional clusters with information about the number of potential customers and detailed information about there social status.

4. Case Study Football World Cup and public Viewing in Germany

As shown in the theoretical foundations, the challenge for long-distance coach vendors is to cover peak demands as accurately as possible. False predictions about the demand are going to lead to either lost profits or inflated costs. As presented, the popularity of public viewing events has been increased during the last soccer competitions. Millions of people in Germany attended the several public viewing offers. Because of this, and the fact that Germans (as many other Europeans) are confessed sport tourists, we choose the UEFA EURO 2016 in France as our first case study to verify the assumptions in the structural equation model. Case-study research is an appropriate research method. This method is mainly used to analyze complex real life situations and to understand complex phenomena. A single case study can be used, if that specific case represents a critical, extreme or unique case in a theory. The UEFA competitions meets that requirements.

For 2016, the big German long distance coach suppliers offer several routes to France. The first questions is, how the demand especially on routes to France and to large German Cities with public viewing, based on the success of the team, is going to be changed. An analyses of past competitions shows that the number of public viewing visitors varies depending on the success of the German team. To answer this question, visitor surveys on selected public viewings are carried out. The procedure is carried out randomly by cluster sampling. The test persons are going to be asked at rest and wait situations [63]. Previous studies have shown, that especially very young and elderly people prefer coach services. These and other socio-demographic characteristics from the model are defined as selection criteria for the survey. In addition to travel preferences, the behavior in social networks and how these channels are used to discuss planned trips is going to asked.

The second aim of the case study is an analysis about the discussion culture of soccer fans in different social media channels. Additionally, the analysis is focused on the discussed topics and the question if fans uses social networks to discuss their travel intentions. To identify relevant data sources in common, a comprehensive source analysis has to be performed. For this purpose, relevant websites and social media channels are going to be analyzed and selected. This pre-study is focused on social-media channels used during the UEFA World Cup 2014. First impressions about the discussion culture and frequencies can be gained from historical data. It is possible to explore from which areas in Germany visitors frequented Public Viewings. For example, users commented on the official site of the Fan Mile
in Berlin, about their travel intentions. Figure 2 shows such an example. The official fan site announced a fan festival in Berlin to visit, whereby one user commented this post: “We arrive from Niederrhein”. The pre-study is also used for a functional web crawler test.

For Euro 2016, the web crawler should be configured according to the results of the preliminary analysis. The aim is to ascertain on which platforms users exchange in any form and intensity. During Euro 2016, the web crawler will scan the relevant pages and collect information, classifying the contributions. For selected routes, a forecast about capacities, based upon the available data, will be performed. In particular, routes to the Paris and Berlin are going to be focused. Through a comparison between forecasted and actual capacities, the potential of the approach can be represented. Data about actual capacities are going to be provided by partners. The surveys and the results from the data analysis are used to verify the assumptions in the structural equation model. This first case study is also used for to obtain more information about the structure and popularity of data sources. Additionally, the case study provides insights about the machine learning algorithms used. If necessary, adjustments for further case studies can be made.

5. Conclusion and outlook on further research

This paper discusses the usefulness of analyzing different data sources as prediction instrument for long-distance coach vendors. As stated in the previous research chapter, such an approach has not been discussed in the research community before. For companies, a high benefit arises, because at the moment, only past data analysis is used. Sport tourism, in particular public viewing events, are a chance for long-distance coach vendors to obtain additional income. On the other hand, the transportation market, especially in Germany, is high-competitive. Long-distance coach vendors need reliable tools to predict the demand. Inaccurate predictions lead to either lost profits or inflated costs. As presented in the paper, social network analysis combined with data from other sources (e.g. internal planning systems, vendor independent portals, location-based data) offers an opportunity to obtain more information about travel behaviors of potential customers. We suggest, that through a combined analysis, long-distance coach vendors are enabled to improve their predictions for future capacities significantly. The concept is constructed as a technological platform by using existing technologies. A web crawler systematically browses different social media channels. The results are stored in a database. Data mining technologies, e.g. collaborative filtering, clustering and classification, are used for the analytical processing.

To verify the approach a case-study is suggested. The focus of the case study is on the UEFA EURO 2016 in France, because it can be that public viewing is going to be as popular as during the previous competitions. A pre-study analyzes social-media communications during the UEFA World Cup 2014. During the competition in 2016, the web crawler collect information from selected sites. For selected long-distances coach routes a forecast, based upon the collected data, is going to be performed. Through a comparison between forecasted and actual capacities, the potential of the approach can be represented. The UEFA EURO 2016 is the first case study. Further case studies are necessary to verify the concept.

The presented approach is not limited on the long-distance coach service market. The reason for choosing the long-distance coach service market is due to the deregulation in 2013. Travel-intensive events like festivals, public-viewing or fairs offer potential for long-distance coach travels. In general, the growth potential of the market is still considerable, because there are still routes where no regular rail connection exists. Setting up a long-distance coach travelling route needs less infrastructure in comparison to railway or airlines. It is also possible to adopt the idea for transportation forms. To predict future capacities and to define the
There are some limitations which has to be considered. The presented case study is focused on the German long-distance coach travel market. At the moment, other countries has been disregarded. The regulation in the long-distance coach travel market differs from country to country. It is not ensured that the concept can be adopted. But we suggest that the presented approach is useful for long-distance coach vendors in countries which are similar in terms of size, infrastructure and degree of regulation then Germany. For countries were long distances between cities are common, e.g. the U.S. or Australia, the adoption of the concept may not be useful. Another limitation results from the assumption that users of social media channels uses these platforms to discuss about their travel behaviors and that these information can be used to predict future capacities of coach services. It is also possible that essential social media platforms are disregarded. The model, as described in the previous chapters, is a starting point for a deeper understanding about social media communication about mega events and the potential for travel vendors to use that. At the moment there are some limitations. First of all, the content of comments and articles is not part of the analysis. At the beginning, the crawler has to be configured manually. Not all websites and social media channels can be analyzed because of that. The long term goal is to develop self-learning algorithms which are able to determine the accuracy of the prediction. Data about actual demands are available in the long-coach vendors enterprise systems. The results can be used for a better prediction in the future. There are many other influence factors which are responsible for decisions about attending events, like weather or prices of competitors. Because of that, the prediction model has to be developed further.

6. Acknowledgements

The project upon which this publication is based on is funded by the German Federal Ministry of Economics and Technology (BMWi) under the project number: 18722 BG/2.

7. References


[24] B. Ufer, „Emotions and experiences at the Public Viewing : explorative interdisciplinary analysis of a social phenomenon“, (German), Göttingen, 2010


[46] Independent transport commission n.d., “Long Distance Travel in Britain Prospects in a time of uncertainty.”, Available at: http://www.theitc.org.uk/docs/2.pdf, 05.06.2015.


