Handling Missing Values in GPS Surveys Using Survival Analysis

A GPS Case Study of Outdoor Advertising

Michael May
Fraunhofer IAIS
St. Augustin, Germany
michael.may
@iais.fraunhofer.de

Christine Körner
Fraunhofer IAIS
St. Augustin, Germany
christine.koerner
@iais.fraunhofer.de

Dirk Hecker
Fraunhofer IAIS
St. Augustin, Germany
dirk.hecker
@iais.fraunhofer.de

Martial Pasquier
Swiss Graduate School of
Public Administration
Lausanne, Switzerland
martial.pasquier
@idheap.unil.ch

Urs Hofmann
Swiss Poster Research Plus
Zurich, Switzerland
u.hofmann@spr-plus.ch

Felix Mende
Affichage International
Zurich, Switzerland
felix.mende
@affichage.com

ABSTRACT

GPS technology has made it possible to evaluate the performance of outdoor advertising campaigns in an objective manner. Given the GPS trajectories of a sample of test persons over several days, their passages with arbitrary poster campaigns can be calculated. However, inference is complicated by the early dropout of persons. Other than in most demonstrations of spatial data mining algorithms where the structure of the data sample is usually disregarded, poster performance measures such as reach and gross impressions evolve continuously over time and require non-intermittent observations. In this paper, we investigate the applicability of survival analysis to compensate for missing measurement days. We formalize the task of modeling the visit potential of geographic locations based on trajectory data as our variable of interest results from dispersed events in space-time. We perform experiments on the cities of Zurich and Bern simulating different dropout mechanisms and dropout rates and show the adequacy of the applied method. Our modeling technique is at present part of a business solution for the Swiss outdoor advertising branch and serves as pricing basis for the majority of Swiss poster locations.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining, Spatial Databases and GIS*; G.3 [**Probability and Statistics**]: *Survival Analysis*

ADKDD'09, June 28, 2009, Paris, France.

General Terms

Algorithms, Experimentation

Keywords

GPS, missing values, survival analysis, mobility mining, outdoor advertising

1. INTRODUCTION

Outdoor advertising is one of the oldest forms of promotion for goods, services or events. Until today it plays a major role in the advertising landscape. For example, the Swiss outdoor advertising branch generated net sales of 663 million Swiss Franc [13] (about 440 million Euro) in 2007, which make up about 11% of the total advertisement net sales in Switzerland.

Consequently, the pricing of poster sites is a critical business task and must be justified by performance measures. The two predominant indicators for poster performance are gross impressions and reach. They specify the total number of contacts of a population with a given poster campaign and the percentage of population that passes at least one poster of the campaign within a given period of time, respectively. Naturally, the question arises how these measures can be determined in an objective way for a given poster campaign.

In 2003 the two leading Swiss outdoor advertising companies commissioned a pilot study to trace individual mobility using GPS technology. Today, the GPS studies include the largest metropolitan areas in Switzerland and a number of smaller conurbations. In total, the survey includes more than 10.000 participants for a measurement period between 7-10 days (see Figure 1). For evaluation, the trajectories are intersected with the geographic locations of poster sites, resulting in the number of passages per site and individual. The passages are weighted according to visibility criteria such as the angle and speed of passage. Finally, the resulting poster contacts provide the means to evaluate reach and gross impressions of poster campaigns [7].

The evaluation of poster contacts is complicated by the

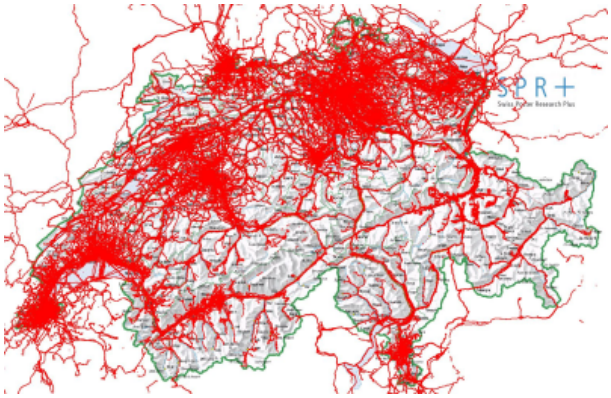

Figure 1: GPS measurements in Switzerland

fact that test persons tend to drop out of the GPS study before the end of the surveying period. As a result, the data set decreases with advancing time. Missing measurement days pose a serious problem to our application as reach and gross impressions are defined with respect to a given time span. As opposed to most case studies of spatial data mining algorithms, the temporal structure of measurements cannot simply be disregarded in our application. If the missing data are ignored, i.e. missing measurement days are treated as immobility, gross impressions and reach will clearly be underestimated. Also, the removal of test persons with less than seven measurement days is not an option as this leads to a strong reduction of available test persons. A third option, typically applied in data mining, is to estimate missing values from the distribution of available measurements. However, this approach is not easily applied to mobility data as it implies the reconstruction of individual trajectories for the missing measurement days. We therefore treat missing data explicitly in the modeling step. In this paper we consider the estimation of reach from incomplete trajectory data using methods from survival analysis. Estimating reach is more challenging than estimating gross impressions because reach depends more strongly on continuous measurements. For evaluation we apply a technique from the area of survival analysis, namely Kaplan-Meier. We show that this approach is adequate to model the reach of poster campaigns for a given audience.

This paper is organized as follows. In the next section we give an overview of related work. We begin with general concepts of missing data and continue with a discussion of related analyses of mobility data. Section 3 provides a general problem statement and formulates the task with respect to outdoor advertisement. Section 4 introduces the applied modeling technique and Section 5 provides experimental results for the conurbations Zurich and Bern. We conclude the paper with a summary and outlook on future work.

2. RELATED WORK

2.1 Concepts of Missing Data

The first major works on missing data appeared in the 1970s. Rubin [9] introduced a typology for missing data and discussed their influence on the inference process. In general, three variants of missing data are distinguished: *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR) [6, 10]. We will start with an intuitive explanation of the concepts given a simple data set and then proceed to a more general formulation. Assume a data set with one explanatory variable X and one dependent variable Y for n objects. While X is completely observed, Y may contain missing values. We can encode the missingness of Y within a separate variable M which assumes a value of 1 if Y is observed and 0 if Y is not observed. Further, we define a variable Z of random noise which is unrelated to X and Y. MCAR occurs if the missingness is completely independent of the data, i.e. $P(M|X,Y) = P(M)$ (see Figure 2a). If a relationship between M and X exists but M is still independent of Y, the data are defined to be MAR. MAR denotes a conditional independence of missingness given a fixed value of X (see Figure 2b). Note however, that under MAR a relationship between M and Y may exist due to their mutual dependency on X. This relationship disappears once the value of X is taken into account. Finally, if the distribution of missing values depends on Y, the data are said to be MNAR (see Figure 2c). MCAR and MAR are also referred to as ignorable missingness (or noninformative dropout in longitudinal studies) while MNAR is termed nonignorable missingness (respectively informative dropout).

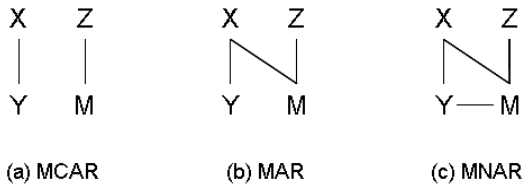

Figure 2: Graphical representation of MCAR, MAR and MNAR (adopted from Schafer and Graham [10])

Usually, data sets contain several variables of which more than one variable may be subject to missingness (e.g. as in longitudinal data). The above illustrated concepts of missing data can then be generalized as follows. Let $Y = (Y_1, Y_2, \ldots, Y_p)$ denote a set of variables with observations for n objects and M denote a $(n \times p)$ matrix encoding (arbitrary) missingness as defined above. Given complete knowledge, we can partition the complete data set $Y_{com} = (Y_{obs}, Y_{mis})$ into subsets containing the values for observed and unobserved parts of the data. The missing values are said to be MCAR if $P(M|Y_{com}) = P(M)$. If the missingness does not depend on the values of unobserved data, i.e. $P(M|Y_{com}) = P(R|Y_{obs})$, the missing values are MAR. Otherwise, the missing data are MNAR [10].

Depending on the type of missingness and the method of inference, estimated parameters of the data may be biased. In general, MCAR results in a correct sampling distribution for Y_{obs} and poses no problem for parameter estimation (except of resulting in reduced sample sizes). Missing values that are MAR produce a correct likelihood distribution, and unbiased parameter estimation conditioned on the observed values is possible. In case of MNAR, parameter estimation is a serious problem and requires an explicit specification of the missingness distribution. However, in many cases the mechanism that leads to missing values in a data set is unknown. The assumption of MAR is then often reasonable

but its robustness should be assured.

2.2 Mobility Analysis

To our best knowledge, the application of survival analysis in order to compensate for missing measurement days in GPS surveys has not been described in literature yet. However, survival analysis has been used by Schönfelder and Axhausen [11] to analyze rhythmic patterns of travel behavior based on travel-diaries. Within the diaries the test persons noted down a categorical purpose of each trip (activity), means of transportation, destination address, time and duration etc. over a period of six weeks [8]. The authors then studied the periodicity of leisure and shopping activities by estimating the belonging survival and hazard functions. Hereby, the activities correspond directly to the events of interest and their geographic location is not considered within the analysis. In contrast, we consider the problem to estimate the visit potential for an arbitrary but fixed set of locations. The number and geographic distribution of the locations play an essential role in our application. For each poster campaign, the passages of a target audience have to be determined explicitly prior to analysis. In this paper we formalize the general task to estimate visit potential of geographic locations and analyze the appropriateness of survival analysis for our application based on a simulation study with the application data.

Fraunhofer IAIS conducts a similar project for the German outdoor advertising media. On behalf of ag.ma, a joint industry committee of German advertising vendors and customers, trajectory data of a nationwide survey are evaluated with methods from survival analysis [2].

3. PROBLEM FORMALIZATION

In most general terms we consider the following problem. Given a set of trajectories Tr of a set of persons P and a set of locations L that may be visited by the persons along their tracks. We are interested in the events of 1^{st}, 2^{nd}, ..., k^{th} passage that the persons produce with the location set over time. We seek aggregated values such as the total number of visits or the percentage of persons that visit the locations $1, 2, \ldots, k$ times within a given time span. Both measures can be derived from the distribution of visits with respect to the time axis.

Note, that this definition is independent of the recording technology and data format of the provided trajectory and location data. Trajectories may be given as raw or preprocessed GPS data, as sequence of radio cells using GSM technology or directly as events recorded at various locations using RFID. Also, the definition of a passage or visit is application dependent and can be specified according to needs.

In the context of our outdoor advertising application, the trajectory and location sets are instantiated as follows. Without loss of generality we assume a discretized geographic space. Personal mobility as well as poster locations are bound to the street network. A street network is a (possibly directed) graph $N = (V, S)$ with nodes $V \subset \mathbb{R}^2$ denoting locations (usually intersections) in 2-dimensional geographic space and edges $S \subseteq V \times V$ denoting street segments. Usually, a street segment $s \in S$ represents a part of road between two intersections, however further division of a real-world street segment may be possible.

The trajectory set $Tr = \{tr_{ij} \mid i = 1..n,\ j = 1..sd\}$ consists of the daily routes tr_{ij} of the test persons p_i $(i = 1..n)$ that have been recorded on day j within the survey duration sd. During data preprocessing the original GPS trajectories have been matched to the street network and are available as finite sequence of traversed street segments $tr_{ij} = (s_1, s_2, \ldots)$. Note, that a trajectory tr_{ij} may be the empty set if a person has stayed at home during a day.

A location set $L \subseteq \mathcal{L}$ represents a specific poster campaign and is a subset of all existing poster locations $\mathcal{L}$. A single poster location $l \in \mathcal{L}$ is thereby modeled by the set of street segments from which the poster can be viewed. A passage of a poster location occurs if the intersection of a trajectory tr_{ij} and a poster location l within a short time span, e.g. 5 minutes, is not empty. The rating of poster locations in outdoor advertising relies further on qualified passages. I.e. a poster contact is not automatically generated by *passing* a poster location but by actually *looking* at a poster [12]. This condition is usually implemented by weighting each passage with visibility criteria such as distance and angle of passage, poster format or illumination at night. However, for simplicity we shall not consider visibility criteria in this paper as they do not affect the general problem setting. Instead, we concentrate on the *coverage* of a campaign, which is a preliminary state of the poster rating index reach.

Definition 1. The coverage of a poster campaign L in a given target group of persons P over a duration of d days is the percentage of persons that produce at least one passage with the poster campaign within the specified time span d.

The evaluation of coverage is based on *first* passages, i.e. a single passage of a person through a location of the poster campaign suffices to increase the coverage. Note, that in the above definition the coverage is defined with respect to the test persons P. Given a representative sample of test persons, the extrapolation to the total population is straightforward as the point estimator of coverage in the population is equal to the coverage in the sample.

Given a complete data set, the coverage for a duration d smaller or equal to the surveying period sd can easily be calculated. However, our data shows that only few people produce measurements for the whole surveying period. For example, in Zurich, about two thirds of the test persons drop out of the study early. For the remaining surveying period the movement behavior of these persons is unknown. The application challenge lies in the appropriate treatment of missing location visits that are caused by incomplete mobility data.

4. MODELING TECHNIQUE

4.1 Data Layout

For a given target audience and campaign our data can be arranged in rectangular form, the rows corresponding to test persons and the columns to observed variables. Some variables $X = (X_1, \ldots, X_p)$ are completely observed such as gender or age. Other variables $Y = (Y_1, \ldots, Y_{sd})$, corresponding to the aggregated number of location visits per measurement day, are available only in part (see Figure 3). The data possess a monotone pattern of nonresponse as for any test person and any $j \in 2..sd$ the following property with respect to Y holds: if Y_j is missing, then $Y_{j+1}, \ldots, Y_{sd}$ are missing as well. A monotone data layout is a special

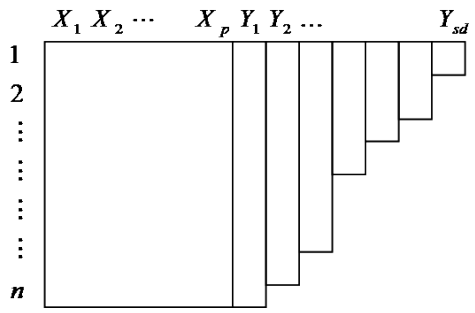

Figure 3: Monotone dropout pattern for GPS test persons

case of a missing data pattern and allows the application of survival analysis with right censoring as explained in detail in the next section.

4.2 Survival Analysis

Survival analysis (also: event history analysis) is a branch of statistics that investigates the occurrence of events as they take place over time. More precisely, survival analysis considers the individual time from an initiating event to an event of interest for a group of objects [1, 5]. Such events denote, for example, the occurrence of some disease in a clinical study or the failure of a device in quality control. One typical method of survival analysis is Kaplan-Meier [4], which estimates the probability that some event does not occur (i.e. the object of interest "survives") within a given period of time allowing for dropout behavior. For example, Kaplan-Meier can be used to calculate the life expectancy after a cancer treatment. Naturally, people enter a medical study at different points in time and therefore possess differing lengths of participation. In addition, people can drop out of the study when moving into another city or dying from a different cause. In our application the event of interest denotes the first passage of a person with a given poster campaign. Dropout occurs if the provided mobility data covers a period of less than seven days. In survival analysis, missing measurements are also termed censored data.

More formally, let T denote a random variable that states the survival time of an object, i.e. the time until the occurrence of the event of interest. The function

$$S(t) = P(T > t) \tag{1}$$

is called the survival function and denotes the probability that the specified event occurs later than some time t. For a given data set, Kaplan-Meier analyzes at which times t_i events occur (with $t_0 = 0$) and determines the following variables

- r_i - number of objects at risk at time t_i,
- e_i - number of events at time t_i,
- c_i - number of dropouts between t_{i-1} and t_i.

In our application the objects at risk at t_0 are all persons in the survey, further $e_0 = c_0 = 0$. At each point in time when events occur, the number of objects at risk is reduced by the objects with events as well as by the objects that drop out in the preceding time interval, i.e. $r_{i+1} = r_i - e_i - c_i$. Kaplan-Meier adapts to differing sample sizes by calculating conditional probabilities between two consecutive events. Objects that drop out of the study between two events are assumed to survive until the next event occurs and are then removed. The conditional probability p_i to survive time point t_i given that t_{i-1} has been survived is calculated as

$$p_i = P(T > t_i | T > t_{i-1}) = \frac{r_{i-1} - e_i}{r_{i-1}}. \tag{2}$$

Given the conditional probabilities p_i, the total probability to survive some time point t_k is

$$S(t_k) = P(T > t_k) = \prod_{i=1}^{k} p_i \tag{3}$$

The transformation from survival probability to poster coverage is straightforward. So far, $S(t)$ states the probability that people in the data sample do not pass any poster location within the campaign until t. Consequently, the coverage of a campaign is given by the probability of the complimentary event

$$F(t) = P(t \leq T) = 1 - S(t). \tag{4}$$

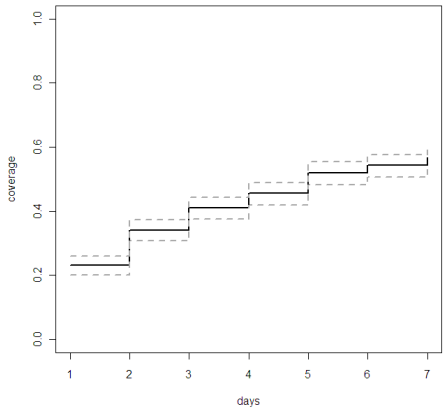

Figure 4: Development of coverage over 7 days

Figure 4 shows the development of coverage for a campaign of 50 randomly selected posters in Zurich over a period of 7 days. Each time a first poster passage occurs an event is generated and the coverage increases. The dotted lines show 95% confidence intervals for the estimated coverage.

4.3 Properties

Kaplan-Meier is a non-parametric estimator and can adapt to arbitrary event rates over time. However, Kaplan-Meier presupposes that the target variable and the dropout behavior are independent of each other, i.e. the method requires MCAR or MAR missingness. (Note, that in case of MAR Kaplan-Meier should be applied to individual strata in order to perform proper conditioning.)

Is MAR a reasonable assumption for our mobility data? Can we presume that dropout behavior and mobility are independent of each other, that mobile and immobile persons

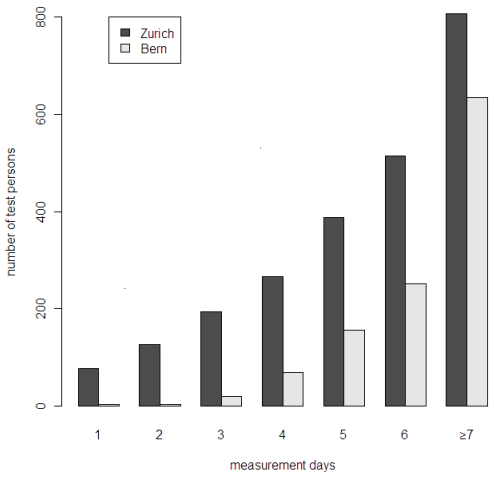

Figure 5: Distribution of valid measurement days per person in Zurich and Bern

are similarly willing to carry GPS devices? At first sight, this supposition seems reasonable. However, differences may arise considering, for example, different age groups. It is known that the mobile behavior of young and old people differs [3]. In addition, young people are usually more technology enthusiastic than old people, which may keep them longer in a survey. Yet, people in their middle years are usually more reliable than young persons.

It is therefore important for our application to analyze the effects of different types of missingness on Kaplan-Meier. How robust will the estimated parameters be in case of violated requirements? What degree of missingness may still produce acceptable results? In the next section we try to answer these questions by simulating different dropout mechanisms and dropout rates in the mobility data.

5. EXPERIMENTS

5.1 Setup

We conducted experiments for the Swiss conurbations Zurich and Bern. Figure 5 shows the distribution of valid measurement days per person for both conurbations. In order to verify our modeling approach, we used only persons with seven valid measurement days and introduced artificial missingness. This resulted in a total of 807 and 635 test persons in Zurich and Bern, respectively. We simulated different dropout behavior and dropout rates and compared the estimated coverage to the coverage in the entire data set. In order to realize different passage probabilities, we experimented with varying campaign sizes (50 and 100 posters in Zurich and 20 and 50 posters in Bern).

We implemented three different dropout strategies, simulating MCAR, MAR and MNAR dropout behavior. In general, we first selected a group of dropout persons according to a given dropout rate and then chose per person a random day (day 2 till 7) from which on all trajectories were censored (i.e. removed). For MCAR, the selection of dropout persons took place completely at random. For MAR, we set up different dropout rates for sociodemographic groups. We used the attributes sex (male, female) and age group (< 30, $30-49$, ≥ 50 years). Within each group, the dropout persons were chosen randomly. Finally, to simulate MNAR we correlated the dropout rate with the total number of passages the persons produced during the surveying period with a given campaign. We simulated positive correlation, censoring preferably persons with many poster passages (mobile persons), and negative correlation, censoring persons with few passages (immobile persons).

We varied the dropout rate between 0.1 and 0.9 (a rate of 0.0 corresponding to the full data set for comparison). For MAR, an increasing dropout rate was implemented for females and the youngest age group, while the dropout rate for the other groups was held constant at 0.2. Note, that in the tables and figures the dropout rate always refers to the varied rate. In case of MAR this does not correspond to the total dropout rate. The results are averages of 100 randomly drawn campaigns of fixed size. For each campaign we performed 10 simulations per dropout behavior to reduce the variability in the results. The calculation of survival rates and confidence intervals was conduced with the statistic software R.

Tables 1-4 contain the estimated coverage for Zurich and Bern. As the results are similar for all four setups, only diagrams for Zurich campaigns of size 50 are shown in Figure 6. The horizontal line denotes the coverage in the full data set. The points denote the mean estimated coverage under different degrees of dropout. The gray lines show 95% confidence intervals calculated based on the cumulative hazard.

5.2 Interpretation

The estimation of coverage for the random dropout strategies MCAR and MAR is unbiased under all tested dropout rates. For MCAR this behavior was expected. For MAR some bias may have been possible due to different mobility within the groups as we conducted all experiments without stratification. In fact, the average coverage between males and females as well as between age groups are slightly different in the data set. For example, in Zurich male test persons produced on average a coverage of 0.712 while female test persons were less mobile and produced a coverage of 0.676 for poster campaigns of size 50. This result implies some degree of robustness of Kaplan-Meier with respect to informative dropout, however this needs to be assessed in further experiments. The diagram for MCAR shows that the confidence intervals increase with advancing dropout rate. This effect is due to the smaller number of test persons that are still at risk at the end of the surveying period. The effect is not visible for MAR because the depicted dropout rate refers only to one gender or age group. For the other groups, dropout was held constant at 0.2, leading to a lower total dropout rate.

For informative censoring the coverage decreases with increasing dropout rate if preferably mobile persons are censored and vice versa for censoring of immobile persons. However, the bias grows more slowly when predominantly mobile persons are censored. This effect is due to the different influence of censoring on mobile and immobile persons. Mobile persons possess a higher chance to produce their first poster passage in the beginning of the surveying period than immobile persons. As the dropout day is chosen randomly once a person has been selected for censoring, the probability that a passage occurs before censoring takes place is

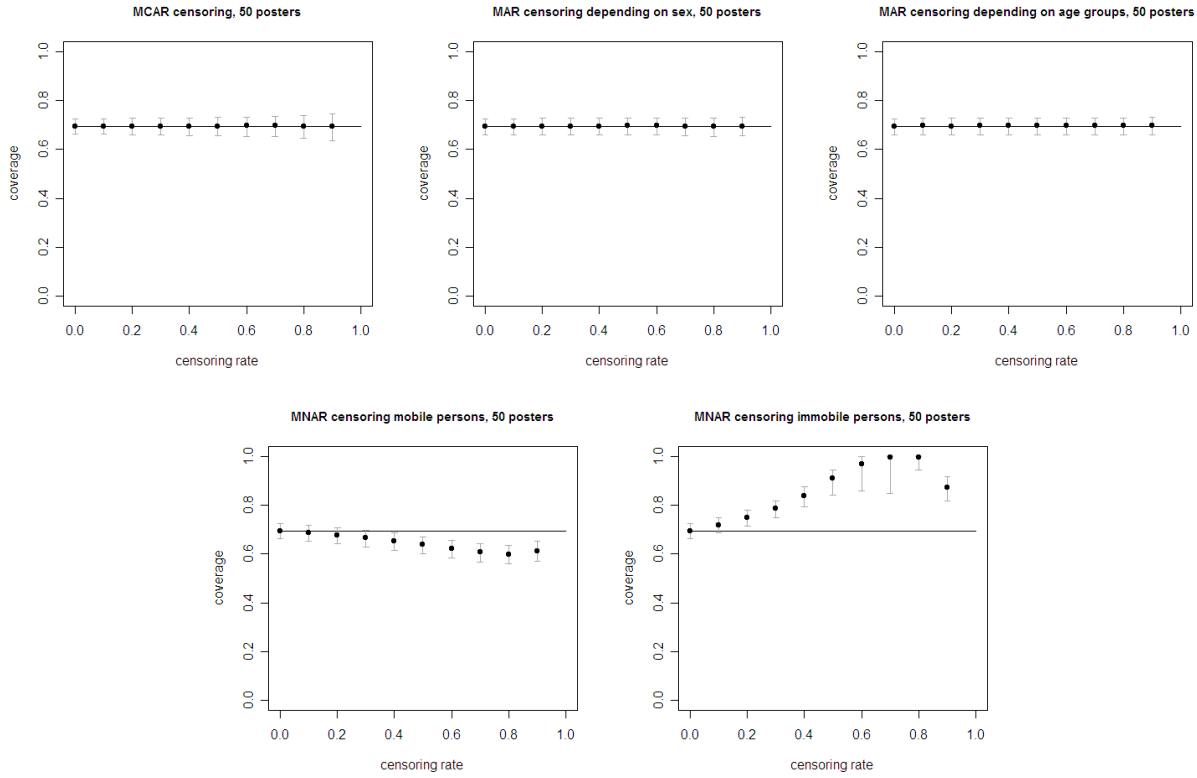

Figure 6: Coverage in Zurich for campaigns of size 50 using the Kaplan-Meier method for MCAR, MAR and MNAR dropout strategies

Table 1: Estimated coverage for campaigns with 50 posters in Zurich

dropout rate	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MCAR	0.696	0.696	0.696	0.696	0.696	0.696	0.696	0.696	0.695	0.695
MAR (sex)	0.696	0.695	0.696	0.696	0.695	0.696	0.696	0.694	0.694	0.696
MAR (age)	0.695	0.696	0.696	0.696	0.696	0.696	0.697	0.697	0.696	0.697
MNAR (mob.)	0.696	0.687	0.677	0.666	0.653	0.638	0.622	0.607	0.599	0.613
MNAR (imm.)	0.696	0.720	0.750	0.786	0.839	0.910	0.970	0.998	0.995	0.871

Table 2: Estimated coverage for campaigns with 100 posters in Zurich

dropout rate	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MCAR	0.832	0.832	0.832	0.832	0.832	0.832	0.832	0.832	0.832	0.833
MAR (sex)	0.832	0.833	0.833	0.832	0.832	0.832	0.831	0.831	0.831	0.831
MAR (age)	0.831	0.832	0.833	0.833	0.833	0.833	0.833	0.833	0.832	0.835
MNAR (mob.)	0.832	0.828	0.822	0.816	0.809	0.800	0.791	0.779	0.764	0.756
MNAR (imm.)	0.832	0.859	0.892	0.933	0.973	0.996	0.999	0.984	0.943	0.846

Table 3: Estimated coverage for campaigns with 20 posters in Bern

dropout rate	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MCAR	0.710	0.710	0.710	0.709	0.710	0.710	0.710	0.710	0.710	0.710
MAR (sex)	0.710	0.709	0.709	0.709	0.709	0.711	0.713	0.712	0.713	0.717
MAR (age)	0.710	0.710	0.711	0.709	0.709	0.710	0.709	0.709	0.708	0.710
MNAR (mob.)	0.710	0.702	0.693	0.682	0.671	0.657	0.643	0.629	0.620	0.631
MNAR (imm.)	0.710	0.734	0.764	0.797	0.849	0.905	0.959	0.985	0.977	0.879

Table 4: Estimated coverage for campaigns with 50 posters in Bern

dropout rate	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
MCAR	0.874	0.874	0.874	0.874	0.874	0.875	0.875	0.874	0.874	0.873
MAR (sex)	0.874	0.875	0.875	0.874	0.875	0.874	0.873	0.874	0.875	0.875
MAR (age)	0.874	0.875	0.874	0.874	0.874	0.874	0.876	0.875	0.875	0.876
MNAR (mob.)	0.874	0.871	0.868	0.863	0.858	0.853	0.845	0.838	0.825	0.813
MNAR (imm.)	0.874	0.901	0.937	0.969	0.991	0.995	0.989	0.974	0.912	0.867

higher for mobile persons than for immobile persons. In consequence, mobile persons are less affected by random censoring of days. This behavior is also reflected in the confidence intervals. While the intervals increase only slightly when censoring preferably mobile persons, censoring immobile persons immediately reduces the persons at risk and results in larger confidence intervals. Both experiment series for MNAR show that the bias starts to decrease at a dropout rate around 0.8. This behavior is natural as the increasing dropout rate lessens the structural effect of deliberate censoring. In the extreme case with a censoring rate of 1.0 all persons would be censored equally.

In summary our experiments confirm the applicability of Kaplan-Meier to mobility data if an uninformative censoring mechanism can be assumed. Further, different censoring rates in distinct sociodemographic groups show no influence on the results even though stratification was not applied. Finally, effects of MNAR censoring depend on the censoring mechanism and seem still acceptable for low censoring rates.

6. CONCLUSION

In this paper we consider the problem to rate arbitrary poster campaigns in outdoor advertising using GPS mobility data that is affected by dropout behavior. We give a formal problem definition and analyze the applicability of survival analysis in an extensive simulation study on a part of the application data. The simulation shows that the proposed method gives unbiased results under systematic censoring in sociodemographic groups even for high censoring rates. Informative censoring leads as expected to biased estimations which, however, may be acceptable for low censoring rates.

In future work we will explore the more general setting of arbitrary patterns of missingness. So far, we have examined a monotone dropout pattern, where test persons quit the study completely. However, persons may also forget to carry the GPS device for a single day within the survey, resulting in intermittent missing values. A second challenge is the estimation of poster ratings for time spans that are longer than the surveying period. In this case, the estimation procedure has to be combined with an appropriate extrapolation model.

7. REFERENCES

[1] O. O. Aalen, Ø. Borgan, and H. K. Gjessing. *Survival and Event History Analysis*. Statistics for Biology and Health. Springer, 2008.

[2] Arbeitsgemeinschaft Media-Analyse e.V. (ag.ma), 2009. http://www.agma-mmc.de.

[3] Bundesministerium für Verkehr, Bau und Stadtentwicklung. Mobilität in Deutschland 2002, Ergebnisbericht (Mobility in Germany 2002, report on results), 2004. http://www.mobilitaet-in-deutschland.de.

[4] E. L. Kaplan and P. Meier. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[5] D. G. Kleinbaum and M. Klein. *Survival Analysis*. Statistics for Biology and Health. Springer, 2005.

[6] R. J. A. Little and D. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability & Mathematical Statistics. John Wiley & Sons, 1987.

[7] M. Pasquier, U. Hofmann, F. H. Mende, M. May, D. Hecker, and C. Körner. Modelling and prospects of the audience measurement for outdoor advertising based on data collection using gps devices (electronic passive measurement system). In *Proceedings of the 8th International Conference on Survey Methods in Transport*, 2008.

[8] PTV AG, B. Fell, S. Schönfelder, and K. Axhausen. Mobidrive questionnaires. Technical report, Institut für Verkehrsplanung, Transporttechnik, Strassen und Eisenbahnbau, ETH Zürich, 2000.

[9] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[10] J. L. Schafer and J. W. Graham. Missing data: Our view on the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

[11] S. Schönfelder and K. Axhausen. Analysing the rhythms of travel using survival analysis. In C. Kaspar, C. Laesser, and T. Bieger, editors, *Jahrbuch 2000/2001 Schweizerische Verkehrswirtschaft*, pages 137–162. Universität St. Gallen, 2001.

[12] J. Z. Sissors and R. B. Baron. *Advertising Media Planning*, chapter 4-5. McGraw-Hill, 2002.

[13] WEMF AG für Werbemedienforschung. Werbeaufwand Schweiz (Advertising expenditure Switzerland). Press release by Stiftung Werbestatistik Schweiz, 2008. http://www.wemf.ch/de/pdf/Pressetext-d.pdf.