PEACOX – Persuasive Advisor for CO2-reducing cross-modal trip planning

Project Reference: 288466

FP7-ICT 2011: 6.6 Low carbon multi-modal mobility and freight transport)

Project Duration: 1 Oct 2011 – 30 Sept 2014

D7.1(1) (updated)

Evaluation Plan

[CURE – Center for Usability Research & Engineering]

Author(s)

Marc Busch (CURE)

Mariella Hager (CURE)

Nadine Rieser (ETHZ)

Lara Montini (ETHZ)

Brian Caulfield (TCD)

Mirjana Artukovic (FLU)

Johann Schrammel (CURE)

Ernst Kruijff (CURE)

Manfred Tscheligi (CURE)

Date: 25/03/2013

Dissemination level: (PU, PP, RE, CO): PU

Project Co-Funded by the European Commission within the 7th Framework Program

Abstract

This deliverable provides a description of the roadmap for the evaluation phases of the developed PEACOX-application-prototype. In addition to the time planning, this deliverable describes the experimental methods used for the evaluation, as well as the selected sample.

Table of Contents

1. Introduction

1.1 Background

This evaluation plan describes the two PEACOX evaluation phases: Evaluation Phase I & Evaluation Phase II. Each of the two evaluation phases will contain: **Lab tests, Field tests and Field trials**. Field tests have a similar procedure to lab tests, but they take place in the actual application context i.e. in the 'field' and not in the lab. Nevertheless, field tests are in a more controlled manner than field trials.

Evaluation phase I contains**: Lab test I, Field test I** and **Field trial I**

Evaluation phase II contains**: Lab test II, Field test II** and **Field trial II**

1.1.1 Scope of this deliverable

This document describes the current state of planning for the evaluation phases of the developed PEACOX-prototype.

2. General overview

This deliverable should give an overview over the planned evaluation phases.

First, we will present the **user evaluation** of the developed prototype (CURE), then, we will present the **evaluation of trip mode and purpose detection functionality** of the prototype (ETZH) and finally, we will present the evaluation of the **emission, exposure and behaviour model** (TCD).

Each of the three evaluation sections will address the following concepts:

- Goals
- Measurements
- Method/procedure
- User involvement
- Timing and resources

Then, the planning of the evaluation phases will be presented.

According to the description of work, the two evaluation phases are planned as following:

Evaluation phase I:

After the first major development phase a first functional prototype system will be tested in Vienna. Targeted number of participating users is 25. The users are encouraged to use their own mobile devices to avoid confounding influences of unfamiliarity with the used device and also to be able to study interaction effects with normal device usage such as making phone calls, writing short messages, etc.

Evaluation phase II:

Towards the end of the project the second system prototype is evaluated with users in Vienna and Dublin.

Targeted number of participating user is 50 (25 in Vienna, 25 in Dublin). The field trials will run for six weeks at the minimum. Basically the same procedure as in the first trial will be applied, which allows for simply comparison of results and evaluation of the progress made.

However, as the system will be applied in different cities there is also an additional focus on evaluating the flexibility and scalability of the system, as the circumstances, data sources, and context is very different in the two cities.

3. CURE: User evaluation

3.1 Goals and state of the art usability and user experience research

The main objective of this evaluation will be placed on the usability and user-experience of the developed system, as well as on changes of environmental attitudes and behaviour.

During the user trials we will not only evaluate the user experience, usability and learnability of the PEACOX-application, but we will also evaluate the effect of the persuasive strategies integrated in the PEACOX-application.

In this section we present an overview and a description of the methods we will use for the PEACOX evaluations. The section is valid for the three parts of the field trials: the lab tests, the field tests and the actual field trials.

3.1.1 Methods used in the lab tests, field tests and field trials

The research to investigate usability, user experience and environmental factors will be based on existing literature in this area.

3.1.1.1 Expert-based evaluation methods

In this sub-section we provide an overview of usability methods, which are used by Human-Computer-Interaction (HCI) experts for evaluating the usability of the system.

3.1.1.1.1 Heuristic evaluations

The goal of a heuristic evaluation[1] is to uncover most usability problems of software without the involvement of end-users. A heuristic evaluation is normally conducted by a handful of experts who are evaluating a system and assigning the found errors to a list of heuristics. These heuristics are rules of the thumb based on long-term HCI experience. HCI experts evaluate if certain software fulfils most of these heuristics. The higher the fulfilment of the heuristics the better the usability of the software is.

[1] A quick and easy user evaluation method based on „rules of thumb“ of good user interface design.

The heuristics we are planning to use are mainly based on several usability heuristics, e.g., Molich and Nielsen (1990) and Mandel (1997) that were extended by several heuristics based on CURE's long-term experience.

The following heuristics for conducting the heuristic evaluations will be used in the PEACOX project:

- **Consistency**: Consistency describes a common design of elements and processes from the users' point of view; all user interface concepts should thus be consistently designed
- **Feedback**: Feedback means that users expect a sufficient system reaction to all of their actions and interactions
- **Efficiency**: The user interface must enable the users to carry out their tasks efficiently
- **Flexibility**: The system must allow different users to work differently, or a single user to work differently if she wishes or needs to, in order to accomplish goals of the users.
- **Clearly marked exits**: The user must always know how to leave a specific context, window or display when working with a user interface, and how the user can return to the starting position
- **Wording in the users' language**: The wording of the user interface must be known and easily understandable to the user
- **Task orientation**: The user interface should always be designed to suit as perfectly as possible the users' tasks; never should a user need to adapt to a system
- **Control**: The user must always be in control of the system; the user must never have the feeling of being controlled by the system
- **Recovery and forgiveness**: The system must prevent the user from (unknowingly) taking severe actions; the user should be able to undo changes or actions easily
- **Minimise memory load**: The user should be able to totally focus on the task, not being troubled with the user interface as such; therefore the user interface must require as little cognitive effort as possible
- **Transparency**: The user must always know what will happen when the user takes an action - the user interface must be transparent

- **Aesthetics and emotional effect**: Everything has an emotional effect; if a user interface has an inappropriate emotional effect, it will interfere with the user's tasks

3.1.1.1.2 Cognitive walkthrough

The cognitive walkthrough method proposed by Wharton et al (1994) is intended to give insights into problems a novice user is expected to have. A usability expert walks through the smartphone-application based on a task-analysis. Based on the task analysis the expert takes different routes through the application with the mind-set of a novice user and analyses every interaction step of the tasks. The focus of the method is "ease of learning", in particular "learning by exploration". Through this method a usability expert is able to uncover, for example, problems in the workflow of an application.

3.1.1.1.3 User groups walkthrough

In user groups walkthroughs the PEACOX user groups [D2.1] are used for an early expert-based evaluation using the cognitive walk-through method. One expert – or a group of experts – steps through a system according to pre-defined context scenarios.

The evaluators imagine themselves to be the user groups and the scenarios are created from the user groups perspective. The single tasks of the scenarios represent the user group's typical interaction with the interface and are selected according to the user groups attributes. Therefore the evaluator is able to see the system through the eyes of the user. In this case, the user is not only one possible and loosely defined person, but resembles a well-defined target group of the system. User group walk-throughs can be conducted in three different steps. Either they are used for a rapid evaluation of a system, which can take about one to two hours, or they can be used for a more formal review with more detailed tasks. Another possibility is to use user groups' walk-throughs as a part of larger design efforts. In PEACOX user group walk-throughs will be used for feedback sessions between HCI researchers and developers. The conduction of this kind of walkthroughs will ensure a strong focus on the needs and wishes of the users of our target groups.

Employing user group walk-throughs, user-typical design issues can be detected early in the design process. Furthermore, also the entire user experience and learnability of a system can

be investigated. The outputs of the user groups based cognitive walkthrough are usability issues, user experience flaws and concerns, but also detailed suggestions for improvement of the system.

The goal of the user group walk-throughs is to not only get insights into the efficiency of a workflow of the application, but rather to focus on the workflow of a defined user.

3.1.1.2 User-based evaluation methods

This sub-section presents evaluation methods in which the participation of end-users is required. User-based evaluation methods complement each other: at the beginning of the development lifecycle expert-based methods ensure a "critical mass" of usability. This prevents expending resources for user-based methods when only low-level prototypes of mock-ups are available. User-based evaluation methods will be used in lab tests, field tests and also in the actual field trials.

3.1.1.2.1 Thinking-Aloud method

Thinking-aloud means that participants have to verbalize their thoughts during a usability evaluation. This method allows the study supervisor to get an insight into possible comprehension problems of the user. Variables such as the task completion rate, occurring interaction problems and errors will be additionally monitored during the evaluations.

3.1.1.2.2 Questionnaires

The usage of questionnaires is very common to conduct research on the users' opinion about a certain product. In this section we present several different types of questionnaires.

Demographic Questionnaire

Demographic questionnaires are used for collecting data from the user such as age, gender or experience with a certain product.

The collected data of this questionnaire allows a comparison between the users' understanding of the evaluated system and her demographical attributes.

Usability Questionnaires

In order to measure the users' perceived usability when using the system we will use several usability measurement methods, such as the **System Usability Scale (SUS).** The SUS was developed by Brooke (1996) and is used for measuring several different aspects of the usability of the evaluated system. The questionnaire consists of ten questions that apply to the evaluated system and which should be rated on a 5-item Likert scale (see **Figure 1**).

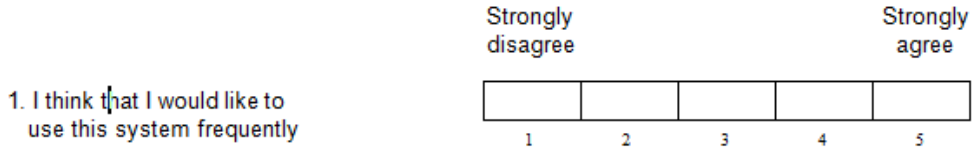

Figure 1: Example of SUS[2]

Environmental attitudes questionnaires

For assessing several aspects of environmental attitudes and behaviour, we will apply following methods: A self-assessment questionnaire with 24 items, the *EAI-24* (Brief version of *Environmental attitudes inventory*; Milfont & Duckitt, 2010), a self-assessment questionnaire with 3 items, the *GreenBehaviorIndex* (World Values Survey, 1999; in Welsch & Kühling,2010), a reduced 6-Item version of the *New Environmental Paradigm (NEP) sc*ale (Dunlap, Van Liere, Mertig & Jones, 2000; Whitmarsh, 2009; in Whitmarsh & O'Neill, 2010) measuring *Pro-environmental values***,** and a 4-item *Pro-environmental self-identity scale* (adapted from Cook et al., 2002; Sparks & Sheperd, 1992; in Whitmarsh & O'Neill, 2010).

[2] © http://www.usabilitynet.org/trump/documents/Suschapt.doc

Emotion questionnaires

The **Self-Assessment Manikin (SAM)** (Bradley & Lang, 1994) is based on pictograms for measuring pleasure, arousal and dominance a user is sensing for a presented product, see Figure 2.

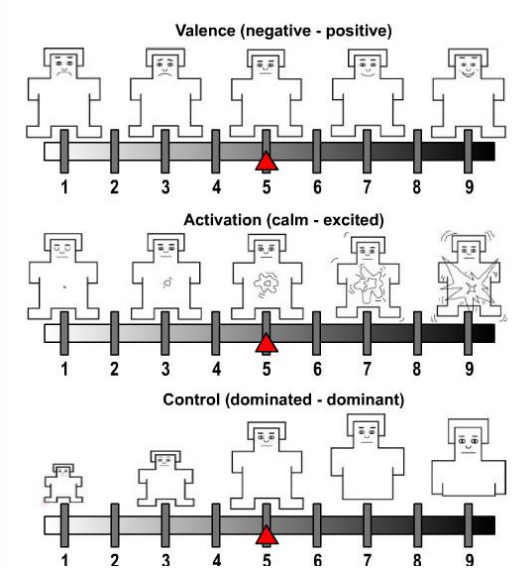

Figure 2: Example of SAM[3]

The **AttrakDiff** is used for measuring the pragmatic as well as the hedonic quality of the software from a users' perspective (Hassenzahl et al., 2003). The results of this questionnaire show how attractive the product is for the users in terms of usability and appearance. The AttrakDiff uses several word-pairs (c.f., Figure 3) for measuring the users' attitudes and emotional state concerning the product.

Figure 3: Example of AttrakDiff[4]

PrEmo (Desmet, 2003) uses 14 explicit emotions illustrated through images. Seven of these emotions are pleasant, the other seven unpleasant (c.f., Figure 4). Users have to rate their feelings towards a product with these 14 illustrated emotions.

[3] © http://2.bp.blogspot.com/_RpwwIGgia4Y/S8-lLUc7rxI/AAAAAAAAAgk/tV7f7kt32q8/s1600/pluginSAM.jpg

[4] © http://www.attrakdiff.de/en/Home/

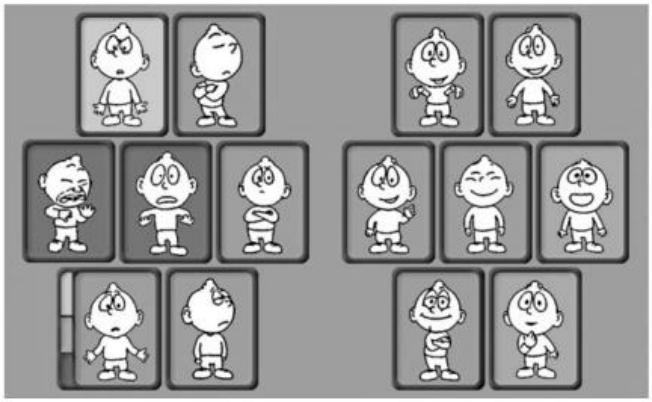

Figure 4: Example of PrEmo

The **EmoCards** (Desmet, 2000) measurement tool measures the users' emotions toward a product through 8 emotion categories represent by EmoCards for each emotion, c.f., Figure 5. The users have to point out the EmoCard that best expresses their most favoured emotional response. EmoCards allow the measurement of Calm - Exited and Pleasant - Unpleasant.

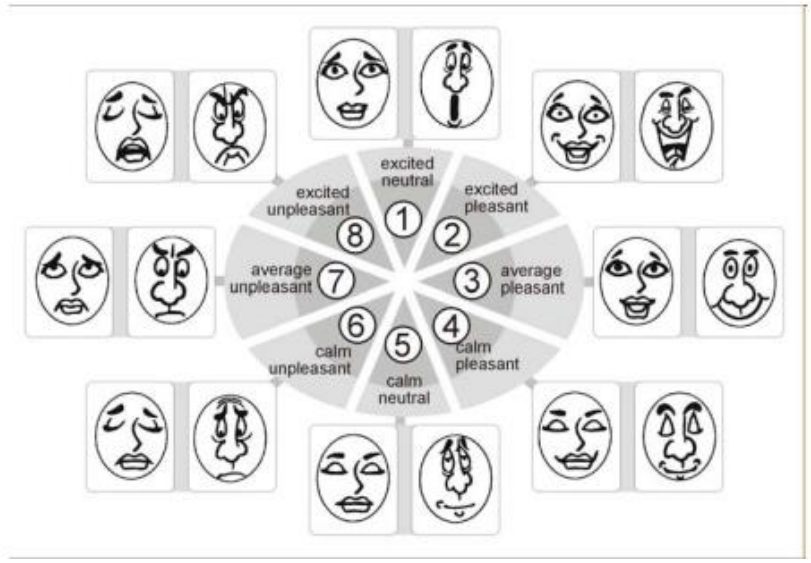

Figure 5: Example of EmoCards

3.1.1.2.3 Interviews and Focus Groups

Interviews are similar to questionnaires, but allow for a more lively discussion between interviewers and participants. Therefore interviews and focus groups allow the collection of more detailed (in-depth) and more complete information concerning the users' ideas, wishes and assumptions. In general it is distinguished between two different kinds of interviews:

Structured Interviews

Structured interviews follow a given questionnaire, which the participant has to answer. The advantage of this method is that it is easier to evaluate and more comparable than the semi-structured interview.

Semi-structured Interviews

Semi-structured interviews give a certain degree of freedom to the interviews. The interviewer has an outline of questions for the interview but the interviewer is allowed to deviate from this outline. This allows an in-depth view into the users' ideas and assumptions but it is harder to evaluate and to compare it to other participants.

3.1.1.2.4 Observations

Users will be observed while interacting with the system. In most cases videos or data are recorded that allow analysing the session's results after its completion. Some systems allow observers to note observations during the actual session.

Observations allow the evaluation supervisor to uncover problems in the handling of an application, e.g., when all participants do not recognize a button from the beginning it can be assumed that the affordance of the button should be improved.

3.2 User involvement

Selected users will be invited to participate in the evaluation activities. In general we plan to conduct the evaluations with 25 persons in the 1st field trial (Vienna) and 50 persons in the

2nd field trial (Vienna and Dublin). This means in total about 75 participants will be involved in the evaluation activities. Trial participants are screened according to the following criteria:

- They must be often involved in personal transportation
- They are interested in new technologies
- They are familiar with mobile devices and have such a device
- Are from diverse demographic backgrounds

CURE will conduct expert and user-based usability evaluations of the systems and setups developed in PEACOX. The user-based evaluations will be usability laboratory evaluations at the user experience labs at CURE. The goal of these evaluations is to measure the usability of the applications in terms of overall user experience, trustworthiness, task efficiency and users satisfaction.

All participants in Vienna will be recruited from CURE's internal database of study participants. This database contains participants with various demographical differences, backgrounds, level of education and more. By selecting a specific subsample of participants (in our case the subsample are the pre-defined user groups) it will be possible to check if there are significant differences in the participants' level of understanding of the evaluated prototypes compared to their demographical background.

All participants in Dublin will be recruited by TCD. This will be done by placing advertisements online calling for participants in the study. Efforts will be made to ensure that the sample is as diverse as possible.

The participants for the PEACOX evaluations will be chosen according to the properties specified by the user groups. This will enable us to evaluate if the PEACOX prototypes matches the requirements of the target group.

CURE also uses an information sheet for informing participants about the evaluations. This sheet has to be signed by the participants before the evaluation starts.

3.3 Timing and resources

Resources of different types need to be available and all sorts of materials need to be prepared to ensure smooth running of the community trials. This section provides an outline of the most important things that have to be considered for the preparation of the trials.

The most important resource is the working PEACOX-prototype. Due to its critical nature fall-back solutions in case of problems should be planned beforehand and stability of the system must be tested thoroughly. The setup of the trials prototype must also include possibilities to log user interactions and provide access to these logs without disturbing the systems functioning.

Laboratories and Meeting rooms at different sites will be required for the conduction of lab tests, focus groups and interviews. Targeted number of participating user is 50 (25 in Vienna, 25 in Ireland).

An environment for experience sampling (triggering of samples, direction towards questionnaires, etc.) and questionnaires will be setup to allow efficient and on-going analysis of data. Also a help desk (e.g. hotline) for users with technical or methodological questions should be established for the duration of the trials. This will be established at CURE in Vienna.

The helpline during the field phases needs to be organised. It is planned to use a mobile phone, which is passed between the different persons responsible for answering the help line.

4. ETZH: Evaluation of trip mode and purpose detection

4.1 Goals and state of the art GPS Processing

In recent years, travel diaries based on person-based GPS observations have become increasingly popular within and beyond the research community. Nowadays, an increasing number of countries use – or consider the usage of – GPS observations in their National Travel Surveys due to their manifold advantages compared to classic survey methods such as paper diaries or telephone interviews. However, there are still several open issues concerning the automated post-processing of these large datasets. Without a reliable post-processing, GPS-based studies require either a considerable amount of manual analysis, leading to costly surveys, or extensive prompted-recall interviews with the respondents. Prompted-recall interviews place a lot of burden on the participants, thus, violating the promise of reducing participant burden made by researchers since the advent of GPS-based travel behaviour studies. Yet, prompted-recall surveys are the only way to establish reliable post-processing routines.

The post-processing routines that have lately been presented by researchers from all over the world are usually organised in sequential modules and contain the following five steps:

- Cleaning and smoothing
- Detection of stages and stop points
- Mode identification
- Activity purpose imputation
- Spatial matching

A sound **cleaning of the data** is essential for meaningful results in the subsequent post-processing steps due to the variety of error sources of GPS measurements. The most commonly used filtering criteria are the number of satellites in view and the PDOP value (e.g. Wolf *et al.*, 1999; Ogle *et al.*, 2002). If these values are not sufficient or available, the stream of GPS points should be scanned for unrealistic position jumps (e.g. Schüssler and Axhausen, 2009). Minor deviations from the true position do not necessarily have to be

filtered but it can help to smooth these positions (e.g. Ogle *et al*., 2002; Chung and Shalaby, 2005; Jun *et al*., 2007; Schüssler and Axhausen, 2009).

The next step is the **detection of stages** and stop points. A stage is a segment of a journey that is covered by one means of transport. Stop points are the time periods in between stages and are either mode transfer points or activities. The stage and stop point detection can either be carried out top-down or bottom-up. Top-down in this context means to start with identifying trips and activities and subsequently breaking the trips down into stages (e.g. Tsui and Shalaby, 2006; Schüssler and Axhausen, 2009) whereas bottom-up approaches first determine stop points and afterwards classify them into activities and transfers (e.g. Moiseeva *et al*., 2010; Marchal *et al*., 2011). Three basic types of stop points can be distinguished: activities with signal loss, activities with ongoing GPS recording and mode transfers. Activities with signal loss are detected by finding time differences between two consecutive GPS points that are longer than a predefined threshold. Activities with ongoing GPS recording result in speeds close to zero (e.g. Schönfelder *et al*., 2006; Tsui and Shalaby, 2006; Schüssler and Axhausen, 2009) or bundles of GPS points (e.g. Doherty et al., 2001; Stopher *et al*., 2005; Schüssler and Axhausen, 2009), i.e. sequences of GPS points positioned very closely to each other. Mode transfers are either characterised by one of the phenomena above or by a change between walking and another mode. These changes can be found using speed and acceleration characteristics of the recorded GPS points (Tsui and Shalaby, 2006; Schüssler and Axhausen, 2009).

Mode detection for person-based GPS can be done with a variety of methods and evaluation criteria. On the one hand, there are rule-based approaches (e.g. de Jong and Mensonides, 2003; Stopher et al., 2005; Chung and Shalaby, 2005; Bohte and Maat, 2008; Marchal *et al*., 2011) that use criteria such as average or maximum speed, duration of the stage, data quality or proximity to certain network elements (e.g. roads, bus stops or train stations) to derive deterministically the best fitting mode. On the other hand there are fuzzy logic approaches (Tsui and Shalaby, 2006; Schüssler and Axhausen, 2009) and Bayesian inference models (Zheng *et al*., 2008; Moiseeva *et al*., 2010) that use similar criteria but account for the fact that many modes have overlapping characteristics, particularly in urban settings, and can therefore only be distinguished with a certain probability.

Relatively few authors have started to work on the **activity purpose** imputation e.g. Wolf et al. (2001), Schönfelder and Samaga (2003), Wolf *et al*. (2004), Stopher *et al*. (2007), Moiseeva *et al*. (2010). All these approaches to derive activity purposes mainly rely on land-use data or locations reported by participants. But land-use data is not available for all areas, the quality is certainly not the same everywhere, and it is uncertain if the databases are always up to date. Another unsolved issue are mixed land-use zones.

The evaluation of processing results is commonly done using **prompted recall** surveys in addition to the GPS diary with automated post-processing. This gives the participants the opportunity to correct and validate the results of the post-processing procedures and to add information that cannot be imputed from the GPS data, e.g. the number of accompanying persons or the scheduling horizon. Moreover, the prompted recall survey delivers the input for learning procedures. A first approach incorporating such **learning procedures** in the imputation of modes and activity purposes was presented by Moiseeva et al. (2010). Regarding the format of the prompted recall survey, the researcher can choose between different options. It can either be conducted as a computer assisted personal interview (CAPI) or as a computer assisted telephone interview (CATI) or as a self-guided web-based interview. The recent trend is towards self-guided web-based prompted recall approaches (e.g. Rieser-Schüssler *et al.*, 2011; Auld *et al*., 2009; Bohte and Maat, 2009; Clark and Doherty, 2010; Giaimo *et al*., 2010). Typically, the data is transmitted via internet or mobile phone communication and the participants are able to review their processed data soon after they uploaded it and at a time most convenient for them. Moreover, the web-based format eases the addition of other survey elements such as stated preference experiments (Oliveira *et al*., 2011) or attitudes and perceptions (Rieser-Schüssler and Axhausen, 2011; Marchal et al., 2011).

The goal of the evaluation with regard to the GPS processing is to assess the accuracy of the improved stage, stop point, mode and trip purpose identification routines. For each algorithm it has to be defined separately how the accuracy is measured as it is not possible to reproduce the travel diary exact to the second. The effect of the inclusion of accelerometer data and learning algorithms will also be evaluated.

4.2 Measurements

Three measurements are used to validate the GPS routines: the accuracy of start and end times of the stages and stop points, the reliability of the mode identified for the stages and the accuracy of the trip purpose identified for the stop points.

4.3 Methods

To validate the output of the GPS routines, the actual stage / stop point times, modes and trip purposes have to be known. It is therefore crucial to first check and correct the data by hand. Ideally, this is done by the participants; a certain amount of validation can also be done by us. However, corrections by us are only possible for start and end times and with high certainty for modes. But for trip purposes the corrections of participants is indispensable. Therefore, a user-friendly prompted recall tool should be used.

To validate **stages and stop points** the routines described for stages in Rieser-Schüssler *et al.* (2011) is used. For start and end times configurable buffers are introduced. This is necessary as even corrected times are not known exact to the second and GPS and accelerometer data can contain gaps of several seconds. A tolerance buffer of 45 seconds around the start and end time of a stage has been found to deliver the most reasonable results.

To get a first impression of how well stages and stop points are detected, the differences of detected and actual stop points together with the distributions of durations of actual and detected stages as well as stop points are used.

For more detailed comparison, each detected stage is assigned to the actual stage during which it took place. More than one detected stage can be assigned to each actual stage if there are some additional stop points wrongly detected. Long and short stages are analysed separately but analogously. Actual stages are then grouped by the number of detected stages they contain, as well as by tolerance buffer criteria. Figure 6 shows the evaluation table that is filled using the described analysis.

Real stage duration	$\lvert\Delta t_{start}\rvert$	$\lvert\Delta t_{end}\rvert$	Assigned detected stages 0	1	2	3	≥4
≤10 min	≤45 s	≤45 s					
		>45 s					
	>45 s	≤45 s					
		>45 s					
>10 min	≤45 s	≤45 s					
		>45 s					
	>45 s	≤45 s					
		>45 s					

Figure 6: Stage evaluation table with an example tolerance buffer of 45 seconds, green are very good matches, red are bad matches and orange matches have to be studied case by case.

To evaluate **mode identification and trip purposes**, stages and stop points are created using the knowledge of the actual stages and stop points. After that the mode and trip purpose identification algorithms are run and the resulting automatically identified modes and trip purposes are compared to the actual modes and trip purposes. The comparison will include the number of mismatches as well as an analysis of which modes and trip purposes are most likely to be confounded with each other.

To evaluate the overall output the user is presented, the modes and trip purposes will also be identified for the stages and stop points detected by our algorithms. These are then compared to the modes and trip purposes of the actual stage and stop point as assigned in the stage evaluation.

To quantify the effect of the processing modules developed for this project, these evaluations will be done with and without accelerometer usage as well as with and without learning routines and usage of user input.

4.4 User involvement

Each field trial user should at least correct one day per week of the travel diary (prompted recall). Ideally, this should be done on a separate prompted recall website. In addition, the app should record all corrections made by the field trial users on the fly. It is especially important that they state their trip purpose as we cannot guess it well enough.

4.5 Timing and resources

The evaluation of the GPS processing methods will start after the field trials. The data has first to be checked / corrected by hand, and then the semi-automated evaluation routines can be used to assess the quality of the routines. 2 PMs are needed for checking.

5. TCD: Evaluation of emission and exposure model

5.1 Goals and state of the art of emission and exposure model

The goal of the evaluation phase of the PEACOX-project is to ensure the model robustness. This process will ensure capability of the models for the intended purpose with satisfactory accuracy, consistent with the objectives (below) of the TCD project component.

Objective 1: Ascertain efficient, accurate and effective methods of estimating CO2 emissions.

Objective 2: Create an emissions model that will predict CO2 emissions from transport before a trip is undertaken.

Objective 3: Create an emissions model that will estimate CO2 emissions from transport in real time or after a trip is complete.

Objective 4: Create a personal exposure model that will provide a simplistic indication of the level of personal exposure.

5.1.1 Methodology: Emissions Models

To calculate and predict emission as accurately (objective 1) as possible in the given mobile device context with existing knowledge on emission factors, the following general methodology (Figure 7) has been developed, which is applicable for both, the real time and the predication model (Objective 2 and 3). To ensure accuracy, the model will account for all possible factors mentioned in the PEACOX Description of Work (DoW). The models has been developed based on the existing emission equations from ARTEMIS Project (Boulter, Barlow, & McCRae, 2009; Boulter & Lathlam, 2009) for a range of private vehicles (e.g. based on fuel type, emission standard, and catalytic converter is included), and emission factors for public transport from an established source (Walsh, Jakeman, Moles, & O'Regan, 2008).

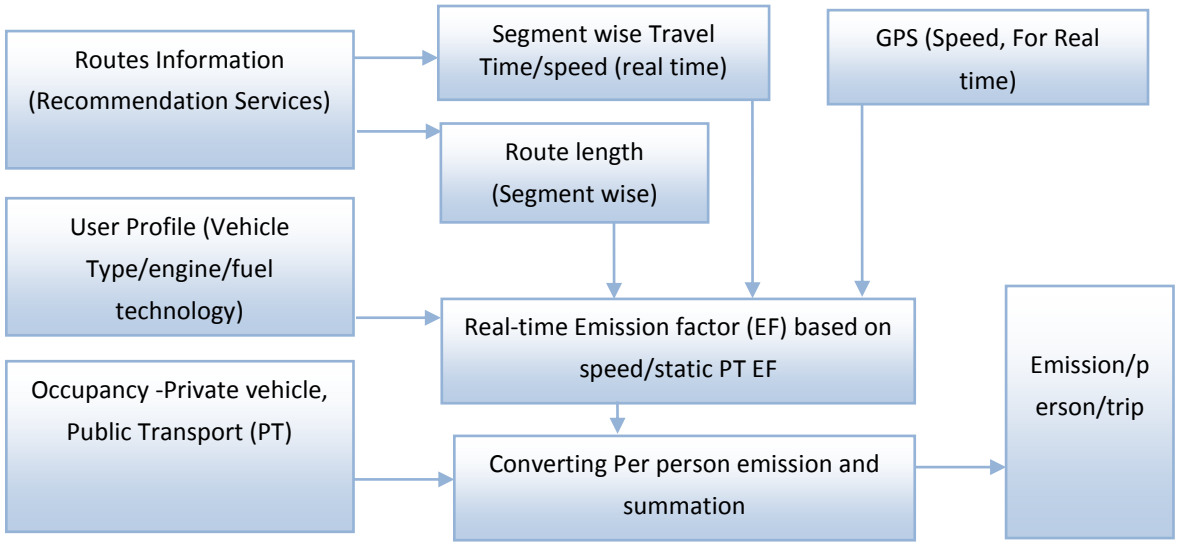

Figure 7: Basic Emission Modelling Methodology

5.1.2 Methodology: Exposure Model

TCD internal database, in addition with some field data will be used to calculate the personal exposure. According to the methodology, two different exposure modelling will be developed in two different levels.

For the first level of analysis, it is necessary to estimate the background concentration of the emission for a particular link. For defining background concentration, land use regression (LUR) can be used among the candidate models, which utilizes the monitored levels of the pollutant of interest as the dependent variable and variables such as traffic, topography, and other geographic variables as the independent variables in a multivariate regression model(Gilliland et al., 2005; Ryan P. H. and LeMasters G. K. 2008. , 2008). After using LUR model, background concentration will be obtained following the idea by(Chen et al., 2010) who used intercept values of Multinomial Linear Regression (MLR) equations for an area's background concentrations.

At the second level analysis, the category of ribbon development along a road will be done by cluster analysis or using a decision tree developed based on the emission dispersion characteristics in a road (e.g. road sides' heavy development, tunnel, etc.). The concept is similar to the much use of various microenvironment (home, other indoor places, transport

and outdoors in(Dons et al., 2011), but extended to the land use planning concept. For every road user, personal exposure could be:

E= Background Concentration+ A1* Volume of Traffic + A2* Ribbon Development Type+ A3* Travel time on the Link+ A4*Wind Flow/Climate; [An= regressing coefficient].

The new set of data in the categorized places will be samples based on the traffic regime say, peak hour, off-peak and moderate flow (e.g. flow 50% capacity) because traffic is the primary and dominant sources of a pollutant. But, some limitation like composition of traffic will not be modelled. To improve the productivity of the model, wind flow and temperature could be considered.

5.5 Measurements

The outcome of the emission model will be mass per kilometre per-person (mass as kilogram for prediction model, and gram for real time model). The output will vary according to the peak and off-peak factor as an occupancy factor was given in the model. In particular, the car emission is also sensitive to vehicle speed, out-side temperature and parking time. These are applicable to both emission prediction model and emission estimation model.

Personal exposure will be the indication of exposure to the level of air pollutant concentrations, expressed as mass per unit volume of atmospheric air (e.g., mg/m3, µg/m3, etc.) in relation to the travel time on a particular link. Although the outcome of the modelling will provide air pollutant concentration, the value will be expressed to the users as a band score. The level of concentration will be given in a scale rating where ‘A’ will indicate excellent travel environment. Similarly, ‘B’ refers ‘Good’, ‘C’ indicates ‘Average’, ‘D’ as ‘Poor’, and ‘E’ refers ‘Unhealthy’ condition.

5.4 Methods

Two types of model require two different types of evaluation. Emission models those are based on the validated ARTEMIS equations and other established emission factors for public transport need model verification only, to ensure the applicability of the emission factors

with sufficient accuracy. Ensuring the functionality of the model in a desired platform will be the priority factor, rather than checking the accuracy of the outcome. Using of an established set of emission factors will also abandon the need for using micro-simulation as the comparative outcome with another emission will be irrelevant for this project. Thus, only functionality checking of the model at the field trial will be justified.

For field testing and validation, a portable emission monitoring system device (PEMS, or a similar device) and mobile GPS device are required. The GPS device will provide speed and acceleration input for cross checking the data by real time emission model whereas, the other device will be required for measuring emission in field tests for validation.

For evaluation of exposure model, tests are necessary for both to check model functionality/verification and model validation. The tests will ensure whether inaccuracies or errors functionality exists in the model.

To validate the model, both sample test and field test is necessary. Field test will also ensure the functionality of the model in a desired platform. On the other hand, certain percentage of dataset will be taken to validate the total model for sample test. Statistical test like the coefficient of determination (R2), or goodness of the fit, etc. will be measured to ensure the calibrated model's efficiency in sample test.

Portable exposure monitoring system device and probably a portable weather monitoring device will be required for the development of exposure regression equations and model validation later.

5.6 User involvement

For functionality test, any vehicle with detail information, and any volunteer will be needed who is able to participate in a multi-modal trip. During a trip, the functionality test for both emission models and exposure model will be carried out.

However, before the functionality test, exposure model field validation is necessary. After validation of the exposure model in the lab test, a field trial may be done to check the accuracy of the model in real world situation in Dublin.

5.7 Resources and Timing

Resources will be needed according to the trial and the lab test. Few devices, vehicles, etc. are the primary requirement for field test. However, before field test, validation of the exposure model will be needed. Thus, involvement of a participant and engagement of few devices will be required early.

6. TCD: Evaluation of behaviour model

6.1 Goals and state of the art of behaviour model

The goal of the behavioural model is to utilize the results of the field trails to create a model which has the potential to predict the mode choices of users of PEACOX-like devices. Many factors both internal and external impact upon the decisions that individuals make every day. While it is impossible to understand the exact workings of an individual's mind when they are under taking a decision, behavioural modelling attempts to explain their choices in terms of factors which the analyst considers important in such a decision making process. These factors are often specific to the choice in question and may vary between individuals. In the case of the PEACOX Project special attention will be given to isolating the impact of the provision of environmental information upon individual's more and route choices.

6.2 Measurements

In terms of measuring the effectiveness of the provision of environmental information, the model will provide outputs such as coefficients of the utility equation (the beta's contained in the utility equation below) for each attribute, such as trip time or cost and, in the case of PEACOX, environmental information. The sign and magnitude of these coefficients will provide a guide to how important each attribute has been in influencing the user's mode choice.

$$V_h = \beta_{1h} + \beta_{2h} X_h$$

Typical Utility Equation

Economic indicators such as Willingness to Pay may also be used if appropriate. In terms of evaluating the effectiveness of the model itself, statistical outputs associated with the Multi Nominal Logit Model such as the Maximum Likelihood Ratio and McFadden's pseudo R squared will be utilized. These indicate the level to which the model describes the factors involved in the users' decision making process. These indicators will then be compared to those already existing in the literature. The goodness of fit of MNL models is very much

dependent upon the industry or sector they are applied to, but typical transport models result in pseudo R squares of 0.25-0.5

Model segmentation techniques will also be used to isolate the influence of socio-economic factors such as gender and age.

$$U_{in} > U_{ij} \forall j \neq i$$

6.3 Methods

The behavioural model will be based upon the Multi Nominal Logit Model and will utilize Random Utility Theory. Random Utility Theory (RUT) is an economic theory which states that an individual derives a certain level of "utility" (U) from each of the alternatives available to him or her in a choice scenario. Utility is a theoretical economic construct of the individual's preferences and is not a measureable quantity. In the case of the PEACOX project a choice scenario is likely to be the choice between transport modes or routes presented to users by the application. According to RUT the individual compares the amount of utility they estimate that they will receive from each alternative and chooses the alternative *i*, which provides the greatest level of utility.

6.4 User Involvement

User involvement should be minimal in terms of evaluating the performance of the model; however user input may be required to explain anomalous observations.

6.5 Resources and Timing

As the behavioural model will be constructed using the observations from the field trials, its evaluation will be dependent on the completion of these trials. An initial model will be constructed using the data from the first Vienna trial and a second model will be produced following the joint trials in Dublin and Vienna. In terms of resources the model will require the appropriate information regarding what mode choices the user made and what information he/she was presented for each trip. In terms of software the behavioural model will be developed using statistical software such as SPSS or Nlogit which will also perform model validation.

7. Detailed planning of evaluation phases

See section 10.1 for a roadmap of the two evaluation phases.

Each of the two **evaluation phases** will include several different interlinked steps.

- **Lab Tests**
- **Field Tests**
- **Field trials**

Evaluation phase I: Lab test I, Field test I, Field trial I

Evaluation phase II: Lab test II, Field test II, Field trial II

The methods used in the evaluation phases are described in sections 3, 4 and 5 and 6.

7.1 Overall preparation of the evaluation phases

Evaluation activities will take place parallel in two tracks in Vienna and Dublin. First, selected users according to the trial needs will be invited to participate in the evaluation activities.

7.2 Implementation of Lab tests (I + II)

As a first part these users will come to the Lab (CUREs usability laboratory in Vienna respectively a room at TCD enhanced with CUREs mobile lab equipment such as recording and screen capturing equipment) and be introduced to the overall procedure and goals of the PEACOX evaluation. Next they participate in the lab test, where the device is explained to them and they are asked to perform several tasks and are observed during the interaction.

7.3 Implementation of Field tests (I + II)

Directly following the lab tests users are driven to nearby sites of the field test and asked to perform a specified set of tasks in this realistic field environment.

Field tests have a similar procedure to lab tests, but they take place in the actual application context i.e. in the 'field' and not in the lab. The first evaluation phase of PEACOX field tests

will take place at two different sites. A central traffic intersection near Vienna and a central traffic intersection near Dublin will be selected for the field tests.

7.4 Implementation of Field trials (I + II)

In the field trials participants are encouraged to use the PEACOX device and application freely and to provide feedback on issues that arise. Trial participants take the device home with them and are free to interact with it as they want. However, to ensure activity several measures are taken by the PEACOX-team to encourage interaction and usage of the system. Since participants cannot be observed as in a lab test, they are asked to take notes, write diaries and fill in protocols. The field trial phase will last for (a minimum of) six weeks.

The following section provides details regarding the planning of the different parts of the two PEACOX-field trials.

The two field trials I and II will be done with split up in two groups: One group (experimental group) is the group with the **PEACOX-prototype** and the other group (control group) will have a **similar application.**

The application that will be used for the control group will fulfil following criteria:

- It will be a routing application
- There will be no (main) focus on environmental awareness or means for the reduction of CO2-consumption

7.4.1 Planning of field trial I

See **Table 1: Outline of the setup of the field trial I in Vienna** for the outline of the setup of field trial I in Vienna.

Vienna	T1	T2	T3
Experimental group: Ecological Aware Navigation-Application	Pre-Assessment – Exposure with prototype	First assessment after 3 weeks exposure	Second assessment after 6 weeks exposure
Control group: Navigation-Application	Pre-Assessment – Exposure with prototype	First assessment after 3 weeks exposure	Second assessment after 6 weeks exposure

Table 1: Outline of the setup of the field trial I in Vienna

7.4.2 Planning of field trial II

See **Table 2: Outline of the setup of the field trial II in Vienna and Dublin** for the outline of the setup for field trial II in Vienna and Dublin.

Vienna	T1	T2	T3	T4
Experimental group: Ecological Aware Navigation-Application	Pre-Assessment – Exposure with prototype	First assessment after 3 weeks exposure	Second assessment after 6 weeks	Third assessment after 9 weeks exposure
Control group: Navigation-Application	Pre-Assessment – Exposure with prototype	First assessment after 3 weeks exposure	Second assessment after 6 weeks exposure	Third assessment after 9 weeks exposure
Dublin	**T1**	**T2**	**T3**	**T4**
Experimental group: Ecological Aware Navigation-Application	Pre-Assessment – Exposure with prototype	First assessment after 3 weeks exposure	Second assessment after 6 weeks	Third assessment after 9 weeks exposure

Control group: Navigation-Application	Pre-Assessment – Exposure with prototype	First assessment after 3 weeks exposure	Second assessment after 6 weeks exposure	Third assessment after 9 weeks exposure

Table 2: Outline of the setup of the field trial II in Vienna and Dublin

8. Conclusion

The present document has provided a detailed overview of the user evaluation phases within PEACOX, as well as the employed methods and the time plan.

As described, the defined methods will be assembled for the different evaluation phases. While the user trials and the virtual reality evaluations will take place within four month phases, the evaluation of developed prototypes will be an on-going activity. Thus, user involvement and a user-centred design process will be assured throughout the project.

For each evaluation phase, an in-detail plan will be developed, and adhere the ethical principles of PEACOX. This plan will include the precise demographics of the participants as well as each document (e.g. questionnaire) in full length that is handed to the user.

9. References

[1] Auld, J., C. Williams, A. K. Mohammadian and P. Nelson (2009) An automated GPS-based prompted recall survey with learning algorithms, Transportation Letters, 1 (1) 59–79.

[2] Bohte, W. and K. Maat (2008) Deriving and validating trip destinations and modes for multiday GPS-based travel surveys: A large-scale application in the Netherlands, paper presented at the 8th International Conference on Survey Methods in Transport, Annecy, May 2008.

[3] Bohte,W. and K. Maat (2009) Deriving and validating trip purposes and travel modes for multiday GPS-based travel surveys: A large-scale application in the Netherlands, Transportation Research Part C: Emerging Technologies, 17 (3) 285–297.

[4] Boulter, P. G., & Lathlam, S. (2009). Emission factors 2009: Report 4-a review of methodologies for modelling cold-start emissions: TRL Limited.

[5] Boulter, P. G., Barlow, T. J., & McCRae, I. S. (2009). Emission factors 2009: Report 3-Exhaust emission factors for road vehicles in United Kingdom: TRL Limited.

[6] Bradley, M.M., Lang. P.J. 1994. Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. Journal of Behavior Therapy and Experimental Psychiatry, 25, 1 (1994).

[7] Brooke, J., 1996. SUS: A quick and dirty usability scale. In: Jordan, P. W., Weerdmeester, B., Thomas, A., Mclelland, I. L. (Eds.), Usability evaluation in industry. Taylor and Francis, London.

[8] Chen, L., Bai, Z., Kong, S., Han, B., You, Y., Ding, X., Liu, A. (2010). A land use regression for predicting NO 2 and PM10 concentrations in different seasons in Tianjin region, China. Journal of Environmental Sciences, 22(9), 1364–1373.

[9] Chung, E.-H. and A. Shalaby (2005) A trip bases reconstruction tool for GPS-based personal travel surveys, Transportation Planning and Technology, 28 (5) 381–401.

[10] Clark, A. F. and S. T. Doherty (2010) A multi-instrumented approach to observing the activity rescheduling decision process, Transportation, 37 (1) 165–181.

[11] de Jong, R. and W. Mensonides (2003) Wearable GPS device as a data collection method for travel research, Working Paper, ITS-WP-03-02, Institute of Transport Studies, University of Sydney, Sydney.

[12] Desmet, P.M.A. 2003. Measuring emotion. In M. Blythe, A Monk, K. Overbeeke, & P. Wright (eds). Funology: From Usability to Enjoyment. Kluwer Academic Press (2003).

[13] Desmet, P.M.A. Emotion through expression: designing mobile telephones with an emotional fit. Report of Modeling the Evaluation Structure of KANSEI, 3 (2000), 103-110.

[14] Doherty, S. T., C. Noel, M. E. H. Lee-Gosselin, C. Sirois, M. Ueno and F. Theberge (2001) Moving beyond observed outcomes: Integrating Global Positioning Systems and interactive computer-based travel behaviour surveys, Transportation Research E-Circular, C026, 449–466.

[15] Dons, E., Panis, L. I., Poppel, Theunis, J., Willems, H., Torfs, R., & Wets, G. (2011). Impact of time activity patterns on personal exposure to black carbon. Atmospheric Environment, 45, 3594-3602.

[16] Giaimo, G., R. Anderson, L.Wargelin and P. R. Stopher (2010)Will it work? Pilot results from the first large-scale GPS-based household travel survey in the United States, Transportation Research Record, 2176, 26–34.

[17] Gilliland, F., Avol, E., Kinney, P., Jerrett, M., Dvonch, T., Lurmann, F., McConnell, R. (2005). Air pollution exposure assessment for epidemiologic studies of pregnant women and children: Lessons learned from the Centers for Children's Environmental Health and Disease Prevention Research. Environ Health Perspect, 113(1447–1454). Retrieved from

[18] Hassenzahl, M., Burmester, M. und Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Ziegler, J. und Szwillus, G. (Hrsg.), Mensch & Computer 2003, Interaktion in Bewegung, S. 187-196. Stuttgart: Teubner.

[19] Jun, J., R. Guensler and J. Ogle (2007) Smoothing methods to minimize impact of Global Positioning System random error on travel distance, speed, and acceleration profile estimates, Transportation Research Record, 1972, 141–150.

[20] Mandel, T. 1997. The Elements of User Interface Design. John Wiley & Sons, Inc., New York, NY, USA.

[21] Marchal, P., J.-L. Madre and S. Yuan (2011) Post-processing procedures for person-based GPS data collected in the French National Travel Survey 2007-2008, paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2011.

[22] Milfont, T.L. & Duckitt, J. (2010). The environmental attitudes inventory: A valid and reliable measure to assess the structure of environmental attitudes. Journal of Environmental Psychology, vol.30, pp. 80-94.

[23] Moiseeva, A., J. Jessurun and H. J. P. Timmermans (2010) Semi-automatic imputation of activity travel diaries using GPS-traces, prompted recall and context-sensitive learning algorithms, Transportation Research Record, 2183, 60–68.

[24] Molich, R., and Nielsen, J. (1990). Improving a human-computer dialogue, Communications of the ACM 33, 3 (March), 338-348.

[25] Ogle, J., R. Guensler, W. Bachman, M. Koutsak and J. Wolf (2002) Accuracy of Global Positioning System for determining driver performance parameters, Transportation Research Record, 1818, 12–24.

[26] Oliveira, M., P. Vovsha, J. Wolf, Y. Birotker, D. Givon and J. Paasche (2011) GPS-assisted prompted recall household travel survey to support development of advanced travel model in Jerusalem, Israel, paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2011.

[27] PEACOX Deliverable 2.1.: Description of User Groups and Travelling Context, 2012.

[28] Rieser-Schüssler, N. and K. W. Axhausen (2011) Combining GPS travel diaries with psychometric scales, paper presented at the 9th International Conference on Survey Methods in Transport, Termas de Puyehue, November 2011.

[29] Rieser-Schüssler, N., L. Montini and C. Dobler (2011) Improving post-processing routines for GPS observations using prompted-recall data, paper presented at the 9th International Conference on Survey Methods in Transport, Termas de Puyehue, November 2011.

[30] Ryan P. H. and LeMasters G. K. 2008. , available at (2008). A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233947/

[31] Schönfelder, S. and U. Samaga (2003) Where do you want to go today? - More observations on daily mobility, paper presented at the 3rd Swiss Transport Research Conference, Ascona, March 2003.

[32] Schönfelder, S., H. Li, R. Guensler, J. Ogle and K. W. Axhausen (2006) Analysis of commute Atlanta instrumented vehicle GPS data: Destination choice behavior and activity spaces, paper presented at the 85th Annual Meeting of the Transportation Research Board,Washington, D.C., January 2006.

[33] Schüssler, N. and K. W. Axhausen (2009) Processing GPS raw data without additional information, Transportation Research Record, 2105, 28–36.

[34] Stopher, P. R., Q. Jiang and C. FitzGerald (2005) Processing GPS data from travel surveys, paper presented at the 2nd International Colloqium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications, Toronto, June 2005.

[35] Stopher, P. R., Q. Jiang and C. FitzGerald. (2007). Deducing mode and purpose from GPS data, paper presented at the 11th TRB National Transportation Planning Applications Conference, Daytona Beach.

[36] Tsui, S. Y. A. and A. Shalaby (2006) An enhanced system for link and mode identification for GPS-based personal travel surveys, Transportation Research Record, 1972, 38–45.

[37] Walsh, C., Jakeman, P., Moles, R., & O'Regan, B. (2008). A comparison of carbon dioxide emissions associated with motorised transport modes and cycling in Ireland. Transportation Research Part D(13), 392–399.

[38] Welsch, H. & Kühling, J. (2010). Pro-environmental behavior and rational consumer choice: Evidence from surveys of life satisfaction. Journal of Economic Psychology, 31, 405-420.

[39] Wharton, C., Rieman, J., Lewis, C., Polson, P. 1994: "The cognitive walkthrough method: a practitioner's guide" in J. Nielsen & R. Mack "Usability Inspection Methods" pp. 105-140.

[40] Whitmarsh, L. & O'Neill, S. (2010). Green identity, green living? The role of pro-environmental self-identity in determining consistency across diverse pro-environmental behaviours. Journal of Environmental Psychology, 30, 305-314.

[41] Wolf, J., R. Guensler and W. Bachman (2001) Elimination of the travel diary - experiment to derive trip purpose from Global Positioning System travel data, Transportation Research Record, 1768, 125–134.

[42] Wolf, J., S. Hallmark, M. Oliveira, R. Guensler and W. Sarasua (1999) Accuracy issues with route choice data collection by using Global Positioning System, Transportation Research Record, 1660, 66–74.

[43] Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira and K. W. Axhausen (2004) Eighty weeks of Global Positioning System traces, Transportation Research Record, 1870, 46–54.

[44] Zheng, Y., L. Liu, L. Wang and X. Xie (2008) Learning transportation mode from raw GPS data for geographic applications on the web, paper presented at the 17th World Wide Web Conference, Bejing, April 2008.

10. Appendix

10.1 Roadmap for the two evaluation phases