Seeing Our Signals:
Combining location traces and web-based models for personal discovery

E. Agapie, G. Chen, D. Houston, E. Howard, J. Kim, M. Y. Mun, A. Mondschein, S. Reddy, R. Rosario, J. Ryder, A. Steiner, J. Burke, E. Estrin, M. Hansen, M. Rahimi.

1. Introduction: A new mobility application

Each of us has a complex and reciprocal relationship with our environment. Based on limited knowledge of this interwoven set of influences and consequences, we constantly make choices: where to live, how to go to work, what brands to buy, what to do with our leisure time. These choices evolve into patterns, and these patterns become driving functions of our relationship with the world around us. With increasing ease, devices we carry can sense, process, and transmit data on these patterns for our own use or to share, carefully, with others. In particular, here we will focus on location time series, gathered from GPS-enabled personal mobile devices. From this capacity emerges a new class of hybrid mobile-web applications that, first, enable *personal exploration of our own patterns* and, second, use the same data to *index our life into other available datasets about the world around us*. Such applications, revealing the previously unobservable about our own lives, offer an opportunity to employ mobile technology to illuminate the ramifications of our choices on others and the effects of the "microenvironments" we move through on us.

This paper proposes and demonstrates how easily gathered location time series data can be used as an index into geospatial models to infer personal environmental impact and exposure. It focuses on three areas of interaction between individuals and the environment: transportation mode choice, overall carbon footprint, and opportunities for healthy eating. This class of applications represents a novel use of mobile systems, web-based mapping, and geospatial data and services. They pose interesting technical challenges and require multidisciplinary research. In the spirit of [20] we offer this anecdote:

Glancing at her phone, Lori sees that she has only a few minutes left before her meeting. She pays for her coffee, stows a proofread copy of her daughter's report on a historical article, The Computer for the 21st Century, in her bag, and dashes out. She texts her daughter that she has some ideas for the end, pulls up a traffic report and, sighing, heads off on surface roads, fashionably late to another hour of Powerpoint. Later on, the pair meet downtown for dinner at they heard about on a food blog and head back home together.

Home on a Saturday four days later, they visit their Personal Environmental Impact Report web pages together for fun, comparing results in good-natured banter. Lori tries to use her daughter's low emissions score from the train trip downtown and constant bike riding as another reason to not help her buy a car, but teenage logic sees it as more of a reason for a hybrid. (That way, when she drives instead of her friends, she'll reduce their emissions!) As they're exploring the pages, created dynamically from location data uploaded by their phones, Lori realizes to herself that her surface street route took her right by a school at recess; she decides try a different route and see how her impact score changes next week. She also wonders why they let traffic go so fast there in the first place—if it were a little calmer there might be less drivers tempted to go through there. Looking back to their monthly summaries, the two realize they walk more on the days when they're together. They decide the report on Mark Weiser's article should end with that.

Effective stewardship of our environment requires an understanding of our complex relationship with it that matures through new knowledge and self-reflection. This is true not just for individuals, but on community and global scales as well. Some of the most persistent challenges attributed to urban living occur through a complex set of person-environment interactions. Researchers in urban planning, public health, environmental science, and other fields have increasingly emphasized the importance of disaggregated analysis and modeling, based on individual behavior and detailed environmental information, to better explain phenomena as diverse (but intertwined) as travel, pollution, energy consumption, physical health, and social cohesion. Whether concerned with exposure or impact, the unit of analysis in environmental research and policymaking has increasingly become the individual (in a spatial context) rather than results aggregated by local or regional geographic units.[3,4,7,14,18].

For reasons of cost and convenience, it is rarely feasible for individuals to gather longitudinal data on their own environmental exposure and impact directly. An example is measurement of air pollutant concentrations, which at this time still requires large, expensive equipment to make local measurements of the precision and accuracy needed to improve on coarse-grained aggregate models. However, the availability of geolocated, time-stamped data from a small number of these stationary and mobile sensors is increasing, as is limited-coverage survey data and coarser-grained remote sensing data. We envision these types of data being used in conjunction with environmental models to provide ongoing feedback to individuals based only on their time-location series, with higher accuracy than demographic estimates, and greater feasibility than domain-specific personal instrumentation. The prevalence of positioning support in emerging devices suggests this approach is scalable up to large populations. **It also scales *down* to small numbers of users because it does not require broad dissemination in order for it to provide value to a user**; yet, as more users engage with the system, their participation offers opportunities for model refinement.

The Personal Environmental Impact Report (PEIR) introduced in the story above is our working example of a relevant use case for this technology, and we are developing the components needed for such a system. It echoes the often government-mandated process of creating Environmental Impact Reports that assess and, where necessary, propose mitigation for the biophysical, social, and other important effects of construction and development before major decisions and commitments are made. Our first implementation focuses in particular on (a) exposure to pollution while traveling on urban roadways (air quality), (b) the impact of our own travel-related emissions on locations such as schools and hospitals (air quality/land use planning), (c) access to unhealthy foods (healthy living), and (d) our contribution to greenhouse gas emissions (global warming)

PEIR is a mechanism for longitudinal documentation of both *impact*—what an individual does to the environment—and what the environment does to the individual, or *exposure*. Significant health problems are a result of "complex interactions between genetic and environmental factors", and activity patterns can have a significant influence on personal exposure to environmental risk factors. Individual location traces, combined with activity characterization and micro-environmental models make personal exposure assessment possible in PEIR.

2. System architecture/framework

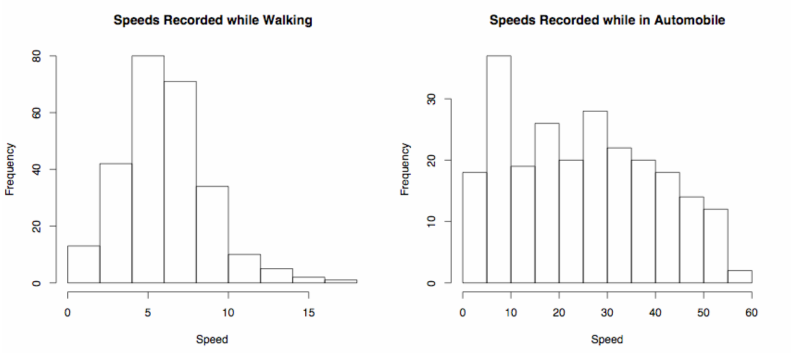
We are developing an online system that combines personal location traces and web based data and models. Our prototype system uses GPS-equipped mobile handsets. We have developed custom handset software (Campaignr, described elsewhere in press) for automatic location time-series collection, robust upload, over-the-air upgrade/tasking, just-in-time annotation with voice or text. We have also begun to develop web server side tools to analyze individual spatiotemporal patterns and calculate corresponding impact and exposure metrics to inform and advise users. Of equal importance are the web-based interfaces informing and advising users, which provide reports, real-time feedback, visualizations and exploratory data analysis tools for non-professional users.

As a PEIR user contributes location traces, feedback on impact and exposure is refined through the following process: (1) **Location trace collection**: Traces of an individual's location are gathered from an installed base of mobile phones using GPS, Cell Tower, and WiFi beaconing. (2) **Trace correction and annotation**: Where possible, the error prone, under-sampled location traces are corrected and annotated using techniques such as map matching with road network and building parcel data. (3) **Activity and location classification**: The corrected and annotated data are automatically classified using web services (e.g., estimating time spent traveling by car vs. on foot for a given day) to provide a first level of refinement to the model output for a given person on a given day. (4) **Context lookup:** Both this corrected, fine-grained location data and the classified data are used as input to web-based information sources on weather, road conditions, real-time traffic monitoring, aggregated driver behaviors, and zoning/planning data. Derived features can reveal significant individual factors such as driver-specific behaviors (e.g., acceleration and braking patterns) (5) **Exposure/Impact Calculation:** Finally, the fine-grained, classified, and derived data are used as input to geospatial data sets and microenvironment models that furthering turn are used to provide an individual's personalized estimates and documentation of localized impacts (e.g., on particular neighborhoods or vulnerable facilities such as schools or retirement homes).

3. Trace Correction and Classification

We can build upon existing work related to activity characterization [1,9,12] and map matching algorithms designed for other contexts to turn a location time series into a geo-coded activity-time series. Using additional map information (such as public transportation stops) the inferred data can indicate not only the velocity of movement, but also with high probability the likelihood that the person is traveling by foot, car, or public transportation.

Figure 1: Speed distributions for walking and automobile



There has been a substantial amount of work on determining activities from location traces. Much of this previous work is based on either machine learning or clustering. Liao et al [12] use relational Markov networks in order to learn high level activities as well as significant locations. By combining raw GPS points with additional features including temporal information, such as time of day and day of week, average speed, and data from geographic information systems; they are able to achieve very high level of accuracy (>90%) for classification of activities. These existing methods have achieved a high amount of success by leveraging historical information about users. In many cases PEIR applications will benefit from this approach.

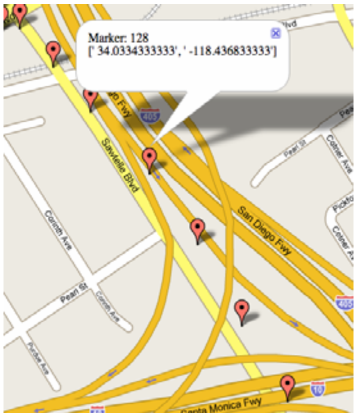
In addition, we are investigating a complementary approach for interpreting activity from real time location traces, in particular, by using the speed attribute in the

GPS trace.

Figure 1 shows the speed distributions of automobile and walking travel modes. The definitive peaks indicate that speed variation has the potential to determine transportation mode. Given a stream of GPS data, we divide the information into a sliding window of two minute chunks. For each chunk, we calculate the max, min, average, and standard deviation of the speed. We create classifiers based on labeled data for the classes of interest (walking, automobile, still) and then evaluate future readings based on these classifiers. Initial results shows that this technique, although simple, is very promising. Based on evaluation on sets of data obtained from a small set of local users with varied transportation patterns, we have achieved over 90% accuracy using Bayesian, linear trees, and support vector machines classifiers. Most of the mis-classifications occurred during transition periods in which a user was going from one state to another (walking to driving an automobile or standing still to taking public transportation). Future work will attempt to estimate speed from Cell Tower traces when GPS is not available. The results will necessarily be far less accurate than with GPS but it will be interesting to see whether the speed estimates will be good enough to support accurate activity classification. Also we plan to consider external map knowledge to infer more specific mode of automobile, public transportation or car, because speed would not be enough to capture that information.

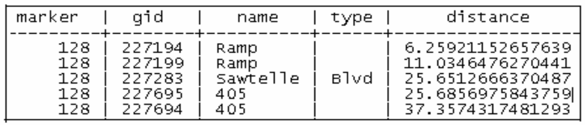
Individual GPS points often deviate from the physical road being traversed due to both inaccuracies in GPS measurements and the maps themselves. Map matching techniques have been developed to improve the interpretation of GPS location data, such as in the case of navigation systems. Similarly, these techniques can be used to improve the performance of activity classification. For example, it would be easier to identify that a person is riding a bus if we find that GPS traces stop along the bus route. Liao et al [12] use conditional random fields to associate GPS measurements with street patches prior to inferring activities.

Figure 2: Inaccurate GPS points



Naive map matching module find the nearest road segment as a correct match. Although they are simple and fast, they are sensitive to the spatial road network and often fails in practice [17]. For example, figure 2 shows one example of GPS data errors and the points that depart from the street. The algorithm identified the Highway 10 to highway 405 interchange ramp as the nearest and correct road while a person was driving on a surface street, Sawtelle Blvd, as shown in Table 1. There are two points from this simple example: that it is strikingly simple to classify activities using very simple techniques, albeit with errors, and there are interesting challenges and directions for improvement. For example, Krumm et al [10,11] take into account time constraints as well as distance to the road: the sequence of road matches must be traversable in the time interval computed from measured timestamps. Najjar et al [16] developed an algorithm dealing with distance, direction and velocity using Belief theory and Kalman filtering.

Table 1: Candidate road matches for GPS point marked 128. Road data from ESRI/Teleatlas StreetPro 2007.



```
 marker  |  gid   |   name    | type |      distance
---------+--------+-----------+------+------------------
     128 | 227194 | Ramp      |      | 6.25921152657639
     128 | 227199 | Ramp      |      | 11.0346476270441
     128 | 227283 | Sawtelle  | Blvd | 25.6512666370487
     128 | 227695 | 405       |      | 25.6856975843759
     128 | 227694 | 405       |      | 37.3574317481293
```

4. Technical Challenges

We have identified several technical challenges including: Context lookup and environmental inference, Selective sharing and privacy, and Effective data presentation.

A. Context Lookup and Inference

The personal exploration of patterns enabled by the data above is on its own valuable, but in this system, it is a means to an end: indexing these patterns of activity into other available datasets to glean insight into our own impact and exposure. PEIR uses GPS location trace data to provide the context necessary to make the models useful/actionable/"real" to people. (1) **Exposure to particulates on highways**: This model will determine the amount of time an individual spends in area that puts them at a higher risk of exposure to particulate matter, with in 200 meters of a road segment with high traffic volumes. Caltrans average annual daily traffic counts are the source for determining these traffic volumes. (2) **Exposure to fast food establishments and advertising:** This model will calculate the number of fast food restaurants that are adjacent to the users transportation corridors. (3) **Impact through local emissions on sensitive populations:** Location data is used to determine the emissions generated from the individuals automobile by the use of the California Air Resources Board's (CARB) emissions model. (4) **Impact through carbon footprint associated with consumption of fossil fuels:** Carbon dioxide is calculated through the same CARB model.

B. Responsibility & Privacy

PEIR uses sensitive personal data, and its systems must be designed to minimize the data released from the user's control to avoid various privacy threats [10,11]. For example, instead of releasing complete time series data to a server, a trusted component could execute the algorithms "close" to the data, exporting only the relevant statistics about activities and exposure, release activity classification without location, or location without time, or employ other forms of spatial and temporal cloaking [6,10,11]. In certain cases, derived features may be more sensitive than raw data: researchers have found that *significant places* like home and workplaces provide very useful context information and can be learned using inference processes [10,11], but releasing this data unnecessarily should be avoidable. We are exploring methods that could use patterns of movement among such places without requiring actual coordinates.

C. Representation & Personalization

Applications like PEIR must support users in connecting sensed information to their daily practices and long-term goals. To this end, analysis, aggregation, and alerting certainly must be configurable but also data and high-level inferences navigable from a number of perspectives. Support (1) understanding of own patterns (2) personal development of experimentation & alternatives (3) goal-setting / coaching (4) comparison, (5) understanding of phenomena / relationship.

5. Conclusions and future work

By combining the power of mobile sensing with web-based geospatial data and models, we can begin to see our own signals, the patterns of our daily life as we interact with the world around us. If systems like the Personal Environmental Impact Report are successful, they will provide actionable feedback to individuals to help them to make responsible choices in the stewardship of their own health and that of others. Used institutionally, such approaches could supplement surveys of activity patterns such as the EPA's Consolidated Human Activity Database [ref] in risk assessment studies, though privacy and data security are even more of a concern than in personal reporting.

Citations

[1] Ashbrook, D. and Starner, T. (2002). "Learning significant locations and predicting user movement with GPS". In Proceedings of ISWC. 101–108.

[2] Austion, S.B. et. al. (2005). "Clustering of Fast-Food Restaurants Around Schools: A Novel 3. Application of Spatial Statistics to the Study of Food Environments". American Journal of Public Health. **95** (9). September.

[3] Axhausen, K. W. and T. Garling (1992). "Activity-Based Approaches to Travel Analysis, Conceptual Frameworks, Models, and Research Problems". Transport Reviews **12**(4): 323-341.

[4] Crane, R.,. R. Crepeau. (1998) "Does Neighborhood Design Influence Travel?: Behavioral Analysis of Travel Diary and GIS Data." Working Paper. University of California Transportation Center, UCTC No 374. January. http://www.uctc.net/papers/374.pdf

[5] Golob, T. F., H. Meurs. (1986) "Biases in response over time in a seven-day travel diary". Transportation **13**: 163-181 (1986)

[6] Gruteser, M., D. Grunwald, (2003). "Anonymous Usage of Location-Based Services through Spatial and Temporal Cloaking," Proceedings of First ACM/USENIX International Conference on Mobile Systems, Applications, and Services (MobiSys), San Francisco, CA, May

[7] Gulliver, J. and D.J. Briggs. (2005) "Time-space modeling of journey-time exposure to traffic related air pollution using GIS". Environmental Research **97**: 10-25.

[8] Guttman, A. (1984). "R-Trees A Dynamic Index Structure for Spatial Searching". ACM

[9] Kang, J. H., Welbourne, W., Stewart, B., and Borriello, G. (2004). Extracting places from traces of locations. In Proceedings of WMASH. 110–118.,.

[10] Krumm, J.(2007). "Inference Attacks on Location Tracks", Fifth International Conference on Pervasive Computing (Pervasive 2007), Toronto, Ontario, Canada

[11]Krumm, J. J. Letchner, E. Horvitz. (2007). "Map Matching with Travel Time Constratins,", SAE 2007 World Congress, April 16-19, , Detroit, MI

[12] Liao, L., Fox, D., and Kautz, H. (2005). "Location-based Activity Recognition, Advances in Neural Information Processing Systems"

[13] Liao, L. D. Fox, and H. Kautz. (2007). "Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields". International Journal of Robotics Research

[14] Miller, H. (2007). "Place-Based versus People-Based Geographic Information Science." Geography Compass **1**(3): 503-535

[15] Midden, C., F. G. Kaiser, L. T. McCalley. (2007) "Technology's Four Roles in Understanding Individuals' Conservation of Natural Resources". J. of Social Issues **63**(1): 155-174.

[16] Najjar, E.L., M.E., Bonnifait, P. (2005) "A Road Matching Method for Precise Vehicle Localization using Kalman Filtering and Belief Theory". Autonomous Robots **19** (2): 173-191.

North, R. J., R.B. Noland, W.Y. Ochieng. J.W. Polak. (2006) "Modelling of particulate matter mass emissions from a light-duty diesel vehicle." Transportation Research Part D. **11**: 344-357.

[17] Quddus, M.A., W. Y.Ochieng, R. B. Noland. (2007) "Current Map Matching Algorithm for Transport Applications: State of the art and Future Research Directions," Transportation Research Part C.

[18]Sultana, S., J. Weber (2007). "Journey-to-work patterns in the age of sprawl: Evidence from two midsize southern metropolitan areas." Professional Geographer **59**(2): 193-208.

[19]Ueno, T., F. Sano, O. Saeki, K. Tsuji. (2006). "Effectiveness of an energy-consumption information system on energy savings in residential houses based on monitored data". Applied Energy **83**: 166-183.

[20] Weiser, M. (1991) The Computer for the 21st Century, Scientific American Special Issue on Communications, Computers, and Networks, September 1991

[21] Wilson, E.J., R. Wilson, K.J. Krizek. (2007) "The implication of school choice on travel behavior and environmental emissions." Transportation Research Part D. **12**: 506-518.

[22] Wood G., M. Newborough. (2007) "Influencing user behaviour with energy information display systems for intelligent homes". International Journal of Energy Research **31**(1): 56