Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Far Out: Predicting Long-Term Human Mobility

Adam Sadilek*
Department of Computer Science
University of Rochester
Rochester, NY 14627
sadilek@cs.rochester.edu

John Krumm
Microsoft Research
One Microsoft Way
Redmond, WA 98052
jckrumm@microsoft.com

Abstract

Much work has been done on predicting where is one going to be in the immediate future, typically within the next hour. By contrast, we address the open problem of predicting human mobility far into the future, a scale of months and years. We propose an efficient nonparametric method that extracts significant and robust patterns in location data, learns their associations with contextual features (such as day of week), and subsequently leverages this information to predict the most likely location at any given time in the future. The entire process is formulated in a principled way as an eigendecomposition problem. Evaluation on a massive dataset with more than 32,000 days worth of GPS data across 703 diverse subjects shows that our model predicts the correct location with high accuracy, even years into the future. This result opens a number of interesting avenues for future research and applications.

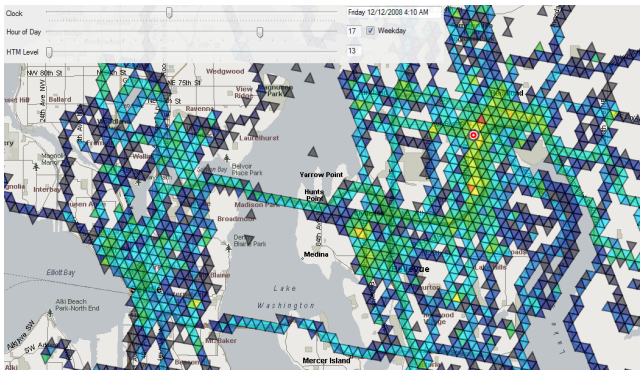


Figure 1: This screenshot of our visualization tool shows mobility patterns of one of our subjects living in the Seattle metropolitan area. The colored triangular cells represent a probability distribution of the person's location given an hour of a day and day type.

Introduction

Where are you going to be 285 days from now at 2PM? This work explores how accurately such questions can be answered across a large sample of people. We propose a novel model of long-term human mobility that extracts significant patterns shaping people's lives, and helps us understand large amounts of data by visualizing the patterns in a meaningful way. But perhaps most importantly, we show that our system, Far Out, *predicts* people's location with high accuracy, even *far into the future*, up to multiple years.

Such predictions have a number of interesting applications at various scales of the target population size. We will give a few examples here. Focusing on one individual at a time, we can provide better reminders, search results, and advertisements by considering all the locations the person is likely to be close to in the future (*e.g.*, "Need a haircut? In 4 days, you will be within 100 meters of a salon that will have a $5 special at that time."). At the social scale (people you know), we can leverage Far Out's predictions to suggest a convenient place and time for everybody to meet, even when they are dispersed throughout the world. We also envision a peer-to-peer package delivery system, but there one would heavily rely on a reliable set of exchange locations, where people are likely to meet in the future. Far Out can provide these. Finally, at the population scale, Far Out is the first step towards bottom-up modeling of the evolution of an entire metropolitan area. By modeling long-term mobility of individuals, emergent patterns, such as traffic congestion, spread of disease, and demand for electricity or other resources, can be predicted a long time ahead as well. These applications motivate the predictive aspect of Far Out, but as we will see, the patterns it finds are also useful for gaining insight into people's activities and detecting unusual behavior. Researchers have recently argued for a comprehensive scientific approach to urban planning, and long-term modeling and prediction of human mobility is certainly an essential element of such a paradigm (Bettencourt and West 2010).

Techniques that work quite well for short-term prediction, such as Markov models and random walk-based models, are of little help for long-term inference. Both classes of models make strong independence assumptions about the domain, and one often postulates that a person's location at time t only depends on her location at time $t-1$. Such models give increasingly poorer and poorer predictions as they are forced to evolve the system further into the future (Musolesi and Mascolo 2009). Although one can improve the performance by conditioning on a larger context and structure the mod-

*Adam performed this work while at Microsoft Research.
Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

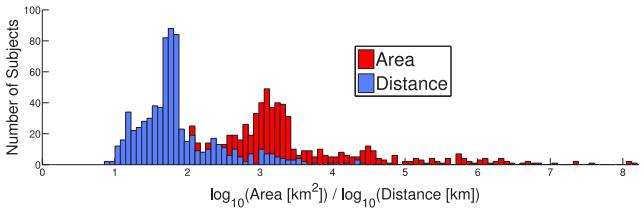


Figure 2: The distribution of the bounding rectangular geographical areas and longest geodesic distances covered by individual subjects.

els hierarchically, learning and inference quickly become intractable or even infeasible due to computational challenges and lack of training data.

While your location in the distant future is in general highly independent of your recent location, as we will see, it is likely to be a good predictor of your location exactly one week from now. Therefore, we view long-term prediction as a process that identifies strong motifs and regularities in subjects' historical data, models their evolution over time, and estimates future locations by projecting the patterns into the future. Far Out implements all three stages of this process.

The Data

We evaluate our models on a large dataset consisting of 703 subjects of two types: people (n 307) and vehicles (n 396). The people include paid and unpaid volunteers who carried consumer-grade GPS loggers while going about their daily lives. Vehicles consist of commercial shuttles, paratransit vans, and personal vehicles of our volunteers, and had the same GPS unit installed on their dashboard. While some of the shuttles follow a relatively fixed schedule, most of them are available on demand and, along with the paratransit vans, flexibly cover the entire Seattle metropolitan area.

Since this work focuses on long-term prediction, we need to consider only datasets that span extensive time periods, which are rare. The number of contiguous days available to us varies across subjects from 7 to 1247 (μ 45.9, σ 117.8). Overall, our dataset contains 32,268 days worth of location data. Fig. 2 shows the distribution of the area (bounding rectangle) covered by our subjects. We observe high variance in the area across subjects, ranging from 30 to more than 10^8 km^2. To put these numbers in perspective, the surface area of the entire earth is 5.2×10^8 km^2.

Methodology and Models

Our models leverage Fourier analysis to find significant periodicities in human mobility, and principal component analysis (PCA) to extract strong meaningful patterns from location data, which are subsequently leveraged for prediction.

To enable Fourier analysis, we represent each GPS reading, consisting of a latitude, longitude pair for each time t, as a complex number z_t latitude$_t$ + (longitude$_t$)i. This allows us to perform Fourier analysis *jointly* over both spatial dimensions of the data, thereby extracting significant periods in a principled way. We can map a function f from time domain to frequency domain via discrete Fourier transform (DFT) given by

$$Z_k \quad \mathscr{F}_t\left[\{z_t\}_{t=0}^{T\ 1}\right](k) \quad \sum_{t=0}^{T\ 1} z_t e^{(\ 2\pi k \frac{t}{T})i} \qquad (1)$$

where $z_0, \ldots z_{T\ 1}$ is a sequence of complex numbers representing a subject's location over T seconds. We refer the reader to (Brigham and Morrow 1967) for more details on DFT.

PCA is a dimensionality reduction technique that transforms the original data into a new basis, where the basis vectors are, in turn, aligned with the directions of the highest remaining variance of the data. PCA can be performed by eigendecomposition of the data covariance matrix, or by applying singular value decomposition (SVD) directly on the data matrix. Our implementation uses the latter approach, as it's more numerically stable. PCA has a probabilistic interpretation as a latent variable model, which endows our model with all the practical advantages stemming from this relationship, such as efficient learning and dealing with missing data (Tipping and Bishop 1999). For a thorough treatment of PCA, see (Jolliffe 2002).

We consider continuous (GPS coordinates) as well as discrete (occupancy grid) data, and our models work with both modalities without *any* modification to the mathematics or the algorithms. In both cases we represent each day as a vector of features. In the continuous representation, we have a 56-element vector shown in Fig. 3. The first 24 elements capture the subject's median latitude for each hour of the day, the next 24 elements correspond to the median longitude, the following 7 elements encode the day of week (in 1-out-of-7 binary code, since it's a categorical variable), and the final element is 1 if the day is a national holiday in the subject's current locale (*e.g.*, Christmas, Thanksgiving) and 0 otherwise. This representation helps us capture the dependence between the subject's location and the hour of the day, day of week, and whether or not the day is a holiday. The continuous representation is best suited for predicting a subject's single, approximate location for a given time, possibly for finding nearby people or points of interest. This representation is not probabilistic, as the discretized representation we describe next.

In the discretized condition, we divide the surface of the globe into equilateral triangular cells of uniform size (side length of 400 meters), and assign each GPS reading to the nearest cell. We then induce an empirical probability distribution over the ten most frequently visited cells and one "other" location that absorbs all GPS readings outside of the top ten cells. Our analysis shows that the 10 discrete locations capture the vast majority of an individual's mobility, and each such cell can often be semantically labeled as home, work, favorite restaurant, *etc.*

Fig. 1 shows the occupancy probability distribution over the cells for one of our subjects, given by

$$\Pr(C \quad c \mid T \quad t, W \quad w) \quad \frac{\text{count}(c,t,w)}{\sum_{c' \in \Omega_C} \text{count}(c',t,w)} \qquad (2)$$

where C, T, and W are random variables representing cells,

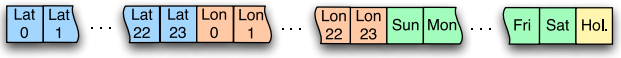


Figure 3: Our continuous vector representation of a day $\boldsymbol{d}$ consists of the median latitude and longitude for each hour of the day (00:00 through 23:59), binary encoding of the day of week, and a binary feature signifying whether a national holiday falls on $\boldsymbol{d}$.

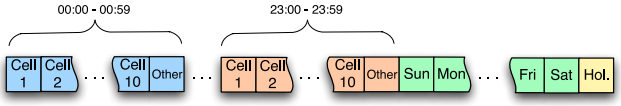


Figure 4: Our cell-based vector representation of a day $\boldsymbol{d}$ encodes the probability distribution over dominant cells conditioned on the time within $\boldsymbol{d}$, and the same day-of-week and holiday information as the continuous representation (last 8 elements).

time of day, and day type, respectively. Ω_C is the set of all cells.

We construct a feature vector for each day from this probability distribution as shown in Fig. 4, where the first 11 elements model the occupancy probability for the 11 discrete places between 00:00 and 00:59 of the day, the next 11 elements capture 01:00 through 01:59, *etc.* The final 8 elements are identical to those in the continuous representation. The discretized representation sacrifices the potential precision of the continuous representation for a richer representation of uncertainty. It does not constrain the subject's location to a single location or cell, but instead represents the fact that the subject could be in one of several cells with some uncertainty for each one.

The decision to divide the data into 24-hour segments is not arbitrary. Applying DFT to the raw GPS data as described above shows that most of the energy is concentrated in periods shorter or equal to 24 hours.

Now we turn our attention to the eigenanalysis of the subjects' location, which provides further insights into the data. Each subject is represented by a matrix $\boldsymbol{D}$, where each row is a day (either in the continuous or the cell form). Prior to computing PCA, we apply Mercator cylindrical projection on the GPS data and normalize each column of observations by subtracting out its mean $\boldsymbol{\mu}$ and dividing by its standard deviation $\boldsymbol{\sigma}$. Normalizing with the mean and standard deviation scales the data so values in each column are in approximately the same range, which in turn prevents any columns from dominating the principal components.

Applying SVD, we effectively find a set of eigenvectors of $\boldsymbol{D}$'s covariance matrix, which we call *eigendays* (Fig. 5).A few top eigendays with the largest eigenvalues induce a subspace, onto which a day can be projected, and that captures most of the variance in the data. For virtually all subjects, ten eigendays are enough to reconstruct their entire location log with more than 90% accuracy. In other words, we can accurately compress an arbitrary day $\boldsymbol{d}$ into only $n \ll |\boldsymbol{d}|$ weights $w_1, \ldots, w_n$ that induce a weighted sum over a common set of ten most dominant eigendays $\boldsymbol{\mathcal{E}}_i$:

$$\boldsymbol{d} \cong \left[\left(\sum_{i=1}^{n} w_i \boldsymbol{\mathcal{E}}_i \right) + \boldsymbol{\mu} \right] \operatorname{diag}(\boldsymbol{\sigma}). \tag{3}$$

This applies to both continuous and discretized data. The reason for this is that human mobility is relatively regular, and there is a large amount of redundancy in the raw representation of people's location. Note that unlike most other approaches, such as Markov models, PCA captures long-term correlations in the data. In our case, this means patterns in location over an entire day, as well as joint correlations among additional attributes (day of week, holiday) and the locations.

Our eigenanalysis shows that there are strong correlations among a subject's latitudes and longitudes over time, and also correlations between other features, such as the day-of-week, and raw location. Let's take eigenday #2 ($\boldsymbol{\mathcal{E}}_2$) in Fig. 5 as an example. From the last 8 elements, we see that PCA automatically grouped holidays, weekends, and Tuesdays within this eigenday. The location pattern for days that fit these criteria is shown in the first 48 elements. In particular, $\boldsymbol{\mathcal{E}}_2$ makes it evident that this person spends her evenings and nights (from 16:00 to 24:00) at a particular constant location in the North-West "corner" of her data, which turns out to be her home.

The last 8 elements of each eigenday can be viewed as indicators that show how strongly the location patterns in the rest of the corresponding eigenday exhibit themselves on a given day-of-week $\times$ holiday combination. For instance, $\boldsymbol{\mathcal{E}}_3$ is dominant on Saturdays, $\boldsymbol{\mathcal{E}}_7$ on Fridays, and $\boldsymbol{\mathcal{E}}_{10}$ on Tuesdays that are not holidays (compare with $\boldsymbol{\mathcal{E}}_2$).

Fig. 6 shows the top ten eigendays for the cell-based representation. Now we see patterns in terms of probability distributions over significant cells. For instance, this subject exhibits a strong "baseline" behavior ($\boldsymbol{\mathcal{E}}_1$) on all days—and especially nonworking days—except for Tuesdays, which are captured in $\boldsymbol{\mathcal{E}}_2$. Note that the complex patterns in cell occupancy as well as the associated day types can be directly read off the eigendays.

Our eigenday decomposition is also useful for detection of anomalous behavior. Given a set of eigendays and their typical weights computed from training data, we can compute how much a new day deviates from the subspace formed by the historical eigendays. The larger the deviation, the more atypical the day is. We leave this opportunity for future work.

So far we have been focusing on the descriptive aspect of our models—what types of patterns they extract and how can we interpret them. Now we turn to the predictive power of Far Out.

Predictive Models

We consider three general types of models for long-term location prediction. Each type works with both continuous (raw GPS) as well as discretized (triangular cells) data, and all our models are directly applied to both types of data without *any* modification of the learning process. Furthermore, while we experiment with two observed features (day

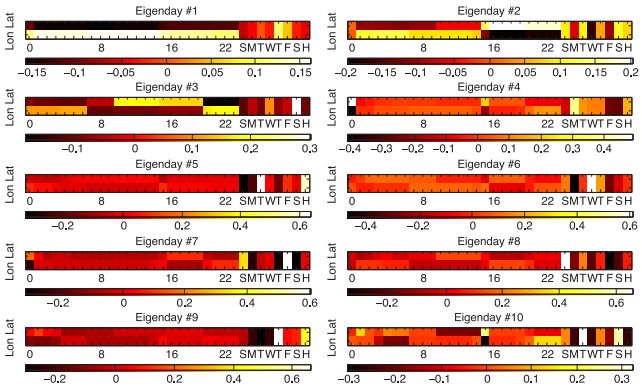


Figure 5: Visualization of the top ten most dominant eigendays ($\boldsymbol{\mathcal{E}}_1$ through $\boldsymbol{\mathcal{E}}_{10}$). The leftmost 48 elements of each eigenday correspond to the latitude and longitude over the 24 hours of a day, latitude plotted in the top rows, longitude in the bottom. The next 7 binary slots capture the seven days of a week, and the last element models holidays versus regular days (*cf.* Fig. 3). The patterns in the GPS as well as the calendar features are color-coded using the mapping shown below each eigenday.

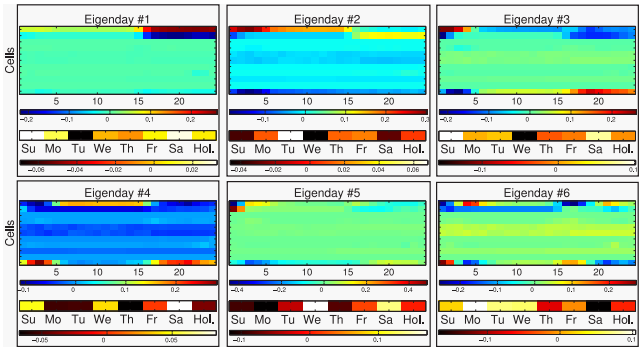


Figure 6: Visualization of the top six most dominant eigendays ($\boldsymbol{\mathcal{E}}_1$ through $\boldsymbol{\mathcal{E}}_6$). The larger matrix within an eigenday shows cell occupancy patterns over the 24 hours of a day. Patterns in the calendar segment of each eigenday are shown below each matrix (*cf.* Fig. 4).

of week and holiday), our models can handle arbitrary number of additional features, such as season, predicted weather, social and political features, known traffic conditions, information extracted from the power spectrum of an individual, and other calendar features (*e.g.*, *Is this the second Thursday of a month?*; *Does a concert or a conference take place?*). In the course of eigendecomposition, Far Out automatically eliminates insignificant and redundant features.

Mean Day Baseline Model For the continuous GPS representation, the baseline model calculates the average latitude and longitude for each hour of day for each day type. In the discrete case, we use the mode of cell IDs instead of the average. To make a prediction for a query with certain observed features o, this model simply retrieves all days that match o from the training data, and outputs their mean or mode. Although simple, this baseline is quite powerful, especially on large datasets such as ours. It virtually eliminates all random noise for repeatedly visited places. Additionally, since the spatial distribution of sporadic and unpredictable trips is largely symmetric over long periods of time, the errors these trips would have caused tend to be averaged out by this model (*e.g.*, a spontaneous trip Seattle-Las Vegas is balanced by an isolated flight Seattle-Alaska).

Projected Eigendays Model First, we learn all principal components (a.k.a. eigendays) from the training data as described above. This results in a $n \times n$ matrix $\boldsymbol{P}$, with eigendays as columns, where n is the dimensionality of the original representation of each day (either 56 or 272).

At testing time, we want to find a fitting vector of weights $\boldsymbol{w}$, such that the observed part of the query can be represented as a weighted sum of the corresponding elements of the principal components in matrix $\boldsymbol{P}$. More specifically, this model predicts a subject's location at a particular time t_q in the future by the following process. First, we extract observed features from t_q, such as which day of week t_q corresponds to.The observed feature values are then written into a query vector $\boldsymbol{q}$. Now we *project* $\boldsymbol{q}$ onto the eigenday space using only the observed elements of the eigendays. This yields a weight for each eigenday, that captures how dominant that eigenday is given the observed feature values:

$$\boldsymbol{w} \quad (\boldsymbol{q} - \boldsymbol{\mu}) \operatorname{diag}(\boldsymbol{\sigma}^{-1}) \boldsymbol{P}_c \tag{4}$$

where $\boldsymbol{q}$ is a row vector of length m (the number of observed elements in the query vector), $\boldsymbol{P}_c$ is a $m \times c$ matrix (c is the number of principal components considered), and $\boldsymbol{w}$ is a row vector of length c. Since we implement PCA in the space of normalized variables, we need to normalize the query vector as well. This is achieved by subtracting the mean $\boldsymbol{\mu}$, and component-wise division by the variance of each column $\boldsymbol{\sigma}$.

Note that finding an optimal set of weights can be viewed as solving (for $\boldsymbol{w}$) a system of linear equations given by

$$\boldsymbol{w}\boldsymbol{P}_c^T \quad (\boldsymbol{q} - \boldsymbol{\mu}) \operatorname{diag}(\boldsymbol{\sigma}^{-1}). \tag{5}$$

However, under most circumstances, such a system is ill-conditioned, which leads to an undesirable numerical sensitivity and subsequently poor results. The system is either over- or under-determined, except when $c \quad m$. Furthermore, $\boldsymbol{P}_c^T$ may be singular.

Theorem 1. *The projected eigendays model learns weights by performing a least-squares fit.*

Proof. If $\boldsymbol{P}$ has linearly independent rows, a generalized inverse (*e.g.*, Moore-Penrose) is given by $\boldsymbol{P}^+ \quad \boldsymbol{P}^*(\boldsymbol{P}\boldsymbol{P}^*)^{-1}$ (Ben-Israel and Greville 2003). In our case, $\boldsymbol{P} \in \mathbb{R}^{m \times c}$ and by definition forms an orthonormal basis. Therefore $\boldsymbol{P}\boldsymbol{P}^*$ is an identity matrix and it follows that $\boldsymbol{P}^+ \quad \boldsymbol{P}^T$. It is known that pseudoinverse provides a least-squares solution to a system of linear equations (Penrose 1956). Thus, equations 4 and 5 are theoretically equivalent, but the earlier formulation is significantly more elegant, efficient, and numerically stable. □

Using Eq. 3, the inferred weights are subsequently used to generate the prediction (either continuous GPS or probability distribution over cells) for time t_q. Note that both training

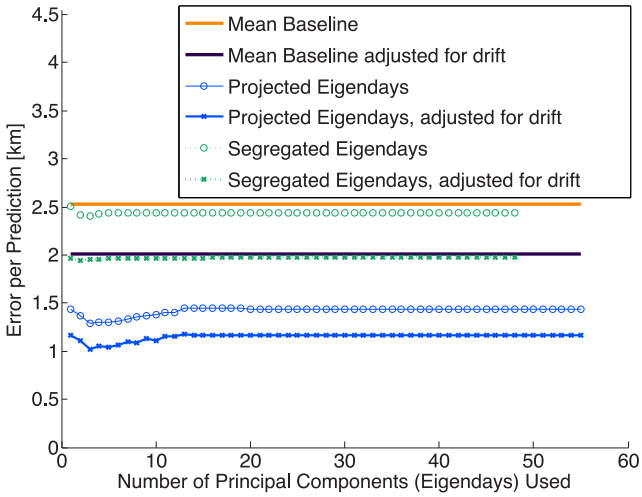


Figure 7: Comparison in terms of absolute prediction error over all subjects as we vary the number of eigendays we leverage.

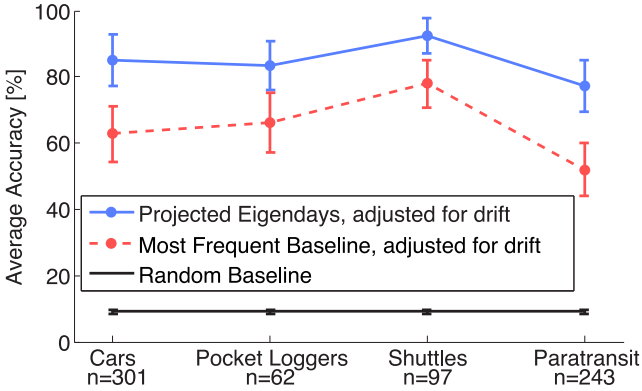


Figure 8: Accuracy of cell-based predictions varies across subject types, but the projected eigendays model outperforms its alternatives by a significant margin.

and testing are efficient ($\mathcal{O}(cdm)$, where d is the number of days) and completely nonparametric, which makes Far Out very easy to apply to other domains with different features.

Segregated Eigendays Model While the last two models induced a single set of eigendays, this model learns a separate library of eigendays for each day type, *e.g.*, eigen-holiday-mondays, over only the location elements of the day vectors $\boldsymbol{d}$. Prediction is made using Eq. 3, where the weights are proportional to the variance each eigenday explains in the training data.

Adapting to Pattern Drift

Since our models operate in a space of normalized variables, we can adapt to the drift of mean and variance of each subject's locations, which does occur over extended periods of time. The basic idea is to weigh more recent training data more heavily than older ones when de-normalizing a prediction (see Eq. 3). We achieve this by imposing a linear decay when learning $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ from the training data.

Experiments and Results

In this section, we evaluate our approach, compare the performance of the proposed models, and discuss insights gained. Unless noted otherwise, for each subject, we always train on the first half of her data (chronologically) and test on the remaining half.

First, let's look at the predictions in the continuous GPS form, where the crucial metric is the median absolute error in distance. Fig. 7 shows the error averaged over all subjects as a function of the number of eigendays leveraged. We show our three model types, both with and without addressing pattern drift. We see that the segregated eigendays model is not significantly better than the baseline. One reason is that it considers each day type in isolation and therefore cannot capture complex motifs spanning multiple days. Additionally, it has to estimate a larger number of parameters than the unified models, which negatively impacts its performance, especially for subjects with smaller amounts of training data.

By considering only the strongest eigendays, we extract the dominant and, in a sense, most dependable patterns, and filter out the volatile, random, and less significant signals. This effect is especially strong in the projected model. Finally, we see that modeling pattern drift systematically reduces the error by approximately 27%.

Now we focus on the evaluation of the same models, but this time they operate on the cell representation. We additionally consider a trivial random baseline that guesses the possible discrete locations uniformly at random. Our eigenday-based models predict based on maximum likelihood:

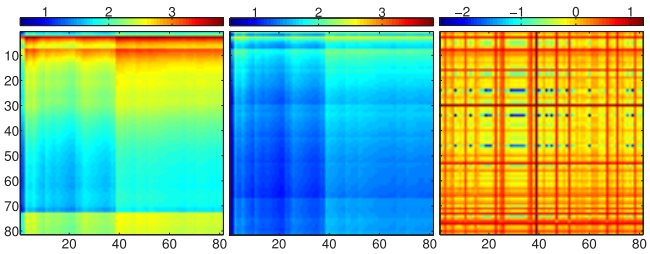
$$c^{\star}_{t,w} \quad \operatorname*{argmax}_{c} \big(\Pr(C \quad c \mid T \quad t, W \quad w)\big).$$

For the sake of brevity, we will focus on the projected eigendays model adapted to pattern drift (with results averaged over c, the number of eigendays used), as our evaluation on the cell-based representation yields the same ordering in model quality as in Fig. 7.

In Fig. 8, we see that the eigenday model clearly dominates both baselines, achieving up to 93% accuracy. Personal cars are about as predictable as pocket loggers (84%), and paratransit vans are significantly harder (77%), as they don't have any fixed schedule nor circadian rhythms.

Since we evaluate on a dataset that encompasses long periods of time, we have a unique opportunity to explore how the test error varies as we make predictions progressively further into the future and increase the amount of training data. Fig. 9 shows these complex relationships for one of our subjects with a total of 162 weeks of recorded data. By contrast, virtually all work to date has concentrated on the first column of *pixels* on the left-hand side of the plot. This is the region of short-term predictions, hours or days into the future.

We see that the projected eigenday model systematically outperforms the baseline and produces a low test error for predictions spanning the entire 81 week testing period (*cf.* Figs. 9a and 9b). In general, as we increase the amount of



(a) Mean Baseline (b) PCA Cumulative (c) PCA Separate

Figure 9: How test error varies depending on how far into the future we predict and how much training data we use. Each plot shows the prediction error, in km, as a function of the amount of training data in weeks (vertical axes), and how many weeks into the future the models predict (horizontal axes). Plots (a) and (b) visualize cumulative error, where a pixel with coordinates (x, y) represents the average error over testing weeks 1 through x, when learning on training weeks 1 through y. Plot (c) shows, on a log scale, the error for each pair of weeks separately, where we train only on week y and test on x.

training data, the error decreases, especially for extremely long-term predictions.

Fig. 9c shows that not all weeks are created equal. There are several unusual and therefore difficult weeks (*e.g.*, test week #38), but in general our approach achieves high accuracy even for predictions 80 weeks into the future. Subsequent work can take advantage of the hindsight afforded by Fig. 9, and eliminate anomalous or confusing time periods (*e.g.*, week #30) from the training set.

Finally, decomposition of the prediction error along day types shows that for human subjects, weekends are most difficult to predict, whereas work days are least entropic. While this is to be expected, we notice a more interesting pattern, where the further away a day is from a nonworking day, the more predictable it is. For instance, Wednesdays in a regular week are the easiest, Fridays and Mondays are harder, and weekends are most difficult. This motif is evident across all human subjects and across a number of metrics, including location entropy, KL divergence and accuracy (cell-based representation), as well as absolute error (continuous data). Shuttles and paratransit exhibit the exact inverse of this pattern.

Related Work

There is ample previous work on building models of short-term mobility, both individual and aggregate, descriptive as well as predictive. But there is a gap in modeling and predicting long-term mobility, which is our contribution (see Table 1).

Recent research has shown that surprisingly rich models of human behavior can be learned from GPS data alone, for example (Ashbrook and Starner 2003; Liao, Fox, and Kautz 2005; Krumm and Horvitz 2006; Ziebart et al. 2008; Sadilek and Kautz 2010). However, previous work focused on making predictions at fixed, and relatively short, time scales. Consequently, questions such as *"Where is Karl going to be in the next hour?"* can often be answered with high accuracy. By contrast, this work explores the predictability of people's mobility at various temporal scales, and specifically far into the future. While short-term prediction is often sufficient for routing in wireless networks, one of the major applications of location modeling to date, long-term modeling is crucial in ubiquitous computing, infrastructure planning, traffic prediction, and other areas, as discussed in the introduction.

	Short Term	**Long Term**
Descriptive	Previous work	Previous work
Predictive	Previous work	**Only Far Out**
Unified	Previous work	**Only Far Out**

Table 1: The context of our contributions.

Much effort on the descriptive models has been motivated by the desire to extract patterns of human mobility, and subsequently leverage them in simulations that accurately mimic observed general statistics of real trajectories (Kim, Kotz, and Kim 2006; González, Hidalgo, and Barabási 2008; Lee et al. 2009; Li et al. 2010; Kim and Kotz 2011). However, all these works focus on aggregate behavior and do not address the problem of location *prediction*, which is the primary focus of this paper.

Virtually all predictive models published to date have addressed only short-term location prediction. Even works with specific long-term focus have considered only predictions up to hours into the future (Scellato et al. 2011). Furthermore, each proposed approach has been specifically tailored for either continuous or discrete data, but not both. For example, (Eagle and Pentland 2009) consider only four discrete locations and make predictions up to 12 hours into the future. By contrast, this paper presents a general model for short- as well as long-term (scale of months and years) prediction, capable of working with *both* types of data representation.

Jeung et al. (2008) evaluate a hybrid location model that invokes two different prediction algorithms, one for queries that are temporally close, and the other for predictions further into the future. However, their approach requires selecting a large number of parameters and metrics. Additionally, Jeung et al. experiment with mostly *synthetic* data. By contrast, we present a unified and nonparametric mobility model and evaluate on an extensive dataset recorded entirely by real-world sensors.

The recent surge of online social networks sparked interest in predicting people's location from their online behavior and interactions (Cho, Myers, and Leskovec 2011; Sadilek, Kautz, and Bigham 2012). However, unlike our work, they address short-term prediction on very sparsely sampled location data, where user location is recorded only when she posts a status update.

In the realm of long-term prediction, (Krumm and Brush 2011) model the probability of being at home at any given hour of a day.We focus on capturing long-term correlations and patterns in the data, and our models handle a large (or even unbounded, in our continuous representation) number

of places, not just one's home.

Conclusions and Future Work

This work is the first to take on understanding and predicting long-term human mobility in a unified way. We show that it is possible to predict location of a wide variety of hundreds of subjects even years into the future and with high accuracy. We propose and evaluate an efficient and nonparametric model based on eigenanalysis, and demonstrate that it systematically outperforms other strong candidates. Since our model operates over continuous, discrete, and probabilistic data representations, it is quite versatile. Additionally, it has a high predictive as well as descriptive power, since the eigendays capture meaningful patterns in subjects' lives. As our final contribution, we analyze the difficulty of location prediction on a continuum from short- to long-term, and show that Far Out's performance is not significantly affected by the temporal distances.

The cell-based modeling is especially amenable to improvements in future work. Namely, since frequently visited cells have a semantic significance, our probabilistic interpretation can be combined in a Bayesian framework with prior probabilities from large-scale surveys[1] and additional constraints, such as physical limits on mobility, where candidate future locations are strictly constrained by one's current position along with means of transportation available. Finally, it would be interesting to generalize the eigenday approach with a hierarchy of nested eigen-periods, where each level captures only the longer patterns the previous one couldn't (*e.g.*, eigendays→eigenweeks→eigenmonths...).

Acknowledgements

We thank Kryštof Hoder, Ashish Kapoor, Tivadar Pápai, and the anonymous reviewers for their helpful comments.

References

Ashbrook, D., and Starner, T. 2003. Using GPS to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Comput.* 7:275–286.

Ben-Israel, A., and Greville, T. 2003. *Generalized inverses: theory and applications*, volume 15. Springer Verlag.

Bettencourt, L., and West, G. 2010. A unified theory of urban living. *Nature* 467(7318):912–913.

Brigham, E., and Morrow, R. 1967. The fast Fourier transform. *Spectrum, IEEE* 4(12):63–70.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and mobility: User movement in location-based social networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Eagle, N., and Pentland, A. 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology* 63(7):1057–1066.

González, M.; Hidalgo, C.; and Barabási, A. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–782.

Jeung, H.; Liu, Q.; Shen, H.; and Zhou, X. 2008. A hybrid prediction model for moving objects. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, 70–79. IEEE.

Jolliffe, I. 2002. Principal component analysis. *Encyclopedia of Statistics in Behavioral Science*.

Kim, M., and Kotz, D. 2011. Identifying unusual days. *Journal of Computing Science and Engineering* 5(1):71–84.

Kim, M.; Kotz, D.; and Kim, S. 2006. Extracting a mobility model from real user traces. In *Proc. IEEE Infocom*, 1–13. Citeseer.

Krumm, J., and Brush, A. 2011. Learning time-based presence probabilities. *Pervasive Computing* 79–96.

Krumm, J., and Horvitz, E. 2006. Predestination: Inferring destinations from partial trajectories. *UbiComp 2006: Ubiquitous Computing* 243–260.

Lee, K.; Hong, S.; Kim, S.; Rhee, I.; and Chong, S. 2009. Slaw: A new mobility model for human walks. In *INFOCOM 2009, IEEE*, 855–863. IEEE.

Li, Z.; Ding, B.; Han, J.; Kays, R.; and Nye, P. 2010. Mining periodic behaviors for moving objects. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1099–1108. ACM.

Liao, L.; Fox, D.; and Kautz, H. 2005. Location-based activity recognition using relational Markov networks. In *IJCAI*.

Musolesi, M., and Mascolo, C. 2009. Mobility models for systems evaluation. a survey.

Penrose, R. 1956. On best approximate solutions of linear matrix equations. In *Proceedings of the Cambridge Philosophical Society*, volume 52, 17–19. Cambridge Univ Press.

Sadilek, A., and Kautz, H. 2010. Recognizing multi-agent activities from GPS data. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Sadilek, A.; Kautz, H.; and Bigham, J. P. 2012. Finding your friends and following them to where you are. In *Fifth ACM International Conference on Web Search and Data Mining*. (Best Paper Award).

Scellato, S.; Musolesi, M.; Mascolo, C.; Latora, V.; and Campbell, A. 2011. Nextplace: A spatio-temporal prediction framework for pervasive systems. *Pervasive Computing* 152–169.

Tipping, M., and Bishop, C. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 611–622.

Ziebart, B.; Maas, A.; Dey, A.; and Bagnell, J. 2008. Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. In *Proceedings of the 10th international conference on Ubiquitous computing*, 322–331. ACM.

[1] *e.g.*, American Time Use Survey, National Household Travel Survey