Learning Likely Locations

John Krumm, Rich Caruana, Scott Counts

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
USA
{jckrumm|rcaruana|counts}@microsoft.com

Abstract. We show that people's travel destinations are predictable based on simple features of their home and destination. Using geotagged Twitter data from over 200,000 people in the U.S., with a median of 10 visits per user, we use machine learning to classify whether or not a person will visit a given location. We find that travel distance is the most important predictive feature. Ignoring distance, using only demographic features pertaining to race, age, income, land area, and household density, we can predict travel destinations with 84% accuracy. We present a careful analysis of the power of individual and grouped demographic features to show which ones have the most predictive impact for where people go.

Keywords: Human mobility, location prediction, Twitter

1 Introduction

We are interested in how we can predict whether or not a person will travel to a given destination away from their home. Such predictions could be used to better understand a person's Web search intentions when they enter an ambiguous place name, and it could be used to target advertising for potential travel destinations. We are also interested in what features of a person's home and candidate destination are important for making accurate predictions. In this paper we examine the predictive power of travel distance, demographics, and spatial features of the home and candidate destinations.

Our training and test data come from geotagged Twitter posts ("tweets") in the U.S. With the introduction of cell phones and mobile social networking sites, researchers have enjoyed a plethora of location data with which to explore human mobility. For example, MIT's Reality Mining dataset offers cell tower data from 100 people over the course of 9 months [1], and has been used for a variety of mobility studies, such as discovering frequent travel patterns [2]. Using data from 100,000 mobile phone users, Gonzalez *et al*. found that people follow simple, predictable patterns and return to just a few locations frequently [3]. In their study of bank note travel, Brockman *et al.* showed that human travel distances decay as a power law [4]. Isaacman *et al.* develop a way to generate synthetic cell phone records and show they can accurately model mobility inside urban areas [5].

Besides cell phones and bank notes, human mobility data exists on many different social networking sites, including Twitter. Certain Twitter client applications let users attach a latitude/longitude to their tweets, and these time-stamped locations give a

sparsely sampled record of the users' travels. Cheng *et al.* present an interesting analysis of over 22 million geotagged tweets from over 225,000 users [6]. They examine the variation in the number geotagged tweets over time, confirm the power law distribution of displacements, and show that residents of different cities have different radii of travel. They also show that the radius of travel generally goes up with increasing income and population density of a person's home region.

As with the work above, we are interested in the interplay between travel propensity, distance, and demographics. However, unlike previous work, we attempt to learn about the attractiveness of a candidate destination, not just how far people will travel. Specifically, we want to compute whether or not a person from a given region will travel to another region, based on the distance between them and the demographic and spatial features of both. This is an important next step in analyzing human mobility, because it begins to answers questions about why people travel, which goes beyond numerical statistics about their behavior. Like Cheng *et al.*, we also use geotagged Twitter data, and we also begin by estimating each user's home location, both of which we describe in the next section.

2 Twitter Data and Home Locations

We gathered geotagged Twitter data for a period of several weeks in mid-2012. Each geotagged tweet comes with a user name, UTC time stamp, and a latitude/longitude giving the user's reported location. After the processing described below, we used data from 213,210 different Twitter users, each with at least one geotagged tweet away from their home.

For each user, we needed an estimate of their home location in order to look up the demographic and spatial features of their home region for later input to our machine learning algorithm. There are many methods for estimating home locations. For instance, Cheng *et al.* uses a recursive search grid to find the mode of the distribution of latitude/longitude locations [6]. Krumm presents four simple home-finding algorithms [7]. The best performing algorithm looked at the latitude/longitude measured nearest to, but not after, 3 a.m. each day and then computed the median latitude and longitude from these points over all days as the home location.

A more principled approach to estimating home location starts by considering when people are normally home. For the U.S., this is available from the American Time Use Survey (ATUS, http://www.bls.gov/tus/). The ATUS is an annual survey sponsored by the U.S. Bureau of Labor Statistics and administered by the U.S. Census Bureau. The survey uses telephone interviews to collect data on how, where, and with whom people spend their time. The raw data is freely available for downloading. We used ATUS data from 2006-2008 inclusive, which consists of survey results from almost 38,000 people. From this, we derived the time-varying probability of a person being at home, averaged into 10-minute bins over a period of one week, shown in Figure 1. This plot shows that the probability of being home peaks at almost 1.0 around 1:30 a.m. on most days and drops to slightly less than 0.4 around noon on

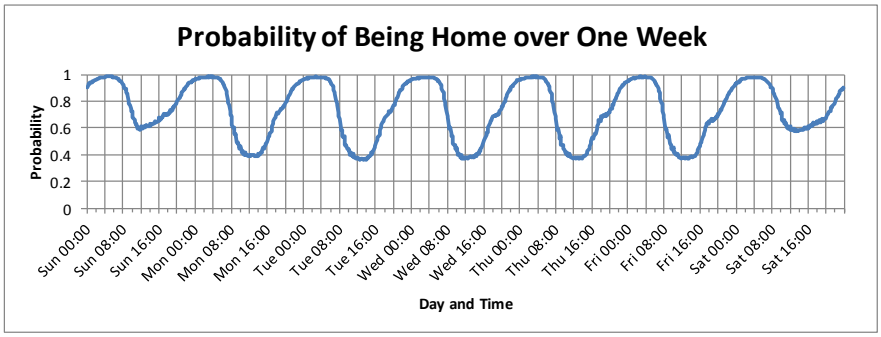

Figure 1: Probability of being home. The probability of being home varies according to the time of day and the day of the week. The data used to derive this function comes from the American Time Use Survey (http://www.bls.gov/tus/). We used this to help estimate a person's home location from time-stamped latitude/longitude values.

most days. The probabilities of being home during the day on weekends are less than for weekdays.

Given a set of time-stamped latitude/longitude values, we would expect the locations occurring at times of larger home probability would be more indicative of a person's home than those occurring at times of lower home probability. We formalize this by taking a weighted median of each Twitter user's latitude/longitude values, where the weights come from the home probabilities in Figure 1. That is, given a time-stamped latitude/longitude, we look up the probability of being home at that time from the data in Figure 1. This probability serves as the weight on both latitude and longitude, and we compute the weighted median of latitude and longitude separately to find the estimate of the home location. We use a median to avoid problems with outliers. Note that Twitter's time stamps are in coordinated universal time (UTC), and ATUS time stamps are in local time. Therefore, we used the latitude/longitude of each geotagged tweet to first convert its time stamp to local time before looking up its corresponding probability of being a home location.

We expect that the user's home location has an effect on where the user will travel. In the next section, we describe the demographic and spatial features we used as input to our predictor.

3 Spatial Features of Visits

For each user in our database, we have their home location, as computed in the previous section, and the locations to which they've traveled, from their geotagged tweets. Our goal is to predict, for a candidate destination, whether or not a person would likely travel there. We make this prediction based on features of the person's home location and the candidate destination. Each geotagged tweet that is not at the per-

son's home serves as a positive example of a visit. This section describes the features we used to characterize these visits for machine learning.

Other research has shown that human travel is governed by distance, with spatial displacements following certain simple distributions [3, 4]. Therefore, one of the features we use is the great circle distance between the user's home and destination. As we will see, this feature is the most powerful predictor of whether or not a user will visit a candidate destination.

The remainder of our features are demographic and spatial features in the regions of the home and destination. We take these from the 2010 American Community Survey (ACS) (http://www.census.gov/acs/www/), which is an ongoing, statistical survey from the U.S. Census Bureau. In contrast to the regular U.S. Census, the ACS is based on a subsample of the U.S. population. Our data gives aggregate demographic and spatial features for each block group in the U.S. A block group is defined by the U.S. Census, and is designed to contain 600 to 3000 people, with an ideal size of 1500 people. Our database contains 217,739 block groups for the U.S. For each block group, we have the fraction that is land (as opposed to water), number of households, median household size, median household income, and fractions of different races and ages. We use these block groups to discretize the map, computing visit probabilities for each distinct pair.

A visit is defined by a move from a user's home block group to a different block group. Approximately 1/3 the features describing a visit come from 20 demographic and spatial features of the block group containing the user's computed home location. For example, one of these features is the median income of the households in the user's home block group. Other features give the fraction of people of difference races and age ranges. We use these home features to account for the possibility that characteristics of a user's home will affect their travel choices. Any tweet occurring outside the user's home block group is considered a visit, and another 1/3 of the visit's features come from the same 20 demographic and spatial features of the destination block group as we use for the home block group, *e.g.* the median income of the destination block group. The destination features account for the possibility that characteristics of the destination will affect the user's travel choices. Thus for each visit we have demographic and spatial features pertaining to the user's home and destination block groups. The last 1/3 of the features come from the differences between the 20 corresponding home and destination features, specifically the value of the destination feature subtracted from the corresponding home feature. For median income, this would be the amount that the home block group's median income exceeds the destination block group's median income. We include these features to account for the possibility that signed differences in features may affect a user's travel choices. For instance, it may be that people tend to visit places with incomes similar to their own.

Thus there are 60 demographic and spatial features for each visit: 20 for the home block group, 20 for the destination block group, and 20 for the differences between the home and destination block groups. In addition to these features, we add one more that gives the great circle distance between the computed home location and the geotagged latitutde/longitude of the destination.

All 61 features are detailed in Table 1. There are many more possible features to compute, such as types and density of businesses, fractions of owned and rented homes, and local attractions. We leave this for future work.

To qualify as a visit, a geotagged tweet must occur in a block group other than the user's home block group. Because of this, we ignore many geotagged tweets in the user's home block group. With this definition of a visit, we have 213,210 users in our study who have made at least one visit. The total number of visits was 3,278,230, with a mean 15.4 per user, and a median 10 per user. The minimum per user was 1, and maximum 21,763.

From our database of geotagged tweets we get positive examples of visits. To learn a classifier, we need negative examples as well. We generate synthetic negative examples for each user by randomly picking block groups that the user was not observed to visit in our data. For balance, we generate the same number of negative examples as we have positive examples for each user. To compute distance to a negative block group, we use the block group's centroid.

4 Classifying Candidate Destinations

Our goal is to create a classifier whose input is a vector of the 61 scalar features described in the previous section and whose output is a probability representing the likelihood that a person in the home block group would visit the destination block group. For training and testing, we have ground truth Twitter data with approximately 3.2 million visits and an equal number of synthetically generated negative examples.

Our classifier is a boosted decision tree. Boosting [8] is a meta-learning procedure that can be applied to different supervised learning methods such as decision trees, neural nets, and SVMs. In this paper we apply boosting to decision trees. In the 1st round of boosting, a decision tree is trained on the data set with equal importance given to all training examples. This tree is added to the model, and the prediction errors it makes on the training set are recorded. In the 2nd round a 2nd tree is trained on the training set which gives more emphasis to the errors made by the 1st model. The predictions of this 2nd tree are added to the predictions of the 1st tree. The

Table 1: These are the groups, names, and descriptions of the 61 features we used for classifying candidate visits. The check marks indicate those features computed for the user's home and destination locations as well as the difference in value between the home and destination locations.

Feature Group	Feature Name	Description	Home	Dest.	Diff.
Distance	Distance	Distance between home and destination			✓
Income	Median income	Median income in U.S. dollars	✓	✓	✓
Race	Non-Hispanic White	Fraction non-Hispanic whites	✓	✓	✓
	Non-Hispanic Black	Fraction non-Hispanic blacks	✓	✓	✓
	Non-Hispanic Indian	Fraction non-Hispanic Indians	✓	✓	✓
	Non-Hispanic Asian	Fraction non-Hispanic Asians	✓	✓	✓
	Non-Hispanic Islander	Fraction non-Hispanic islanders	✓	✓	✓
	Non-Hispanic Other	Fraction non-Hispanic other than above	✓	✓	✓
	Non-Hispanic Two	Fraction non-Hispanic two or more races	✓	✓	✓
	Hispanic	Fraction Hispanic	✓	✓	✓
Age	Under 10	Fraction under 10 years old	✓	✓	✓
	10-19	Fraction 10-19 years old	✓	✓	✓
	20-29	Fraction 20-29 years old	✓	✓	✓
	30-39	Fraction 30-39 years old	✓	✓	✓
	40-49	Fraction 40-49 years old	✓	✓	✓
	50-59	Fraction 50-59 years old	✓	✓	✓
	60-69	Fraction 60-69 years old	✓	✓	✓
	Over 69	Fraction over 69 years old	✓	✓	✓
Household Size	Mean household size	Mean number of persons in household	✓	✓	✓
Household Density	Houshold Density	Households per unit land area	✓	✓	✓

combined predictions of the two trees usually are more accurate than the predictions of either tree alone. This process of adding trees to the prediction model, with each new tree being trained to correct the errors made by the previous trees, is repeated until the model contains a fixed number of trees, or until accuracy stops improving as more trees are added to the model. For the boosted models in this paper, we use 100 iterations of boosting and observed that accuracy improved by less than 2 percentage points after adding more trees up to 1000. In practice, boosted decision trees yield high accuracy models on many problems and for many different metrics [9].

One advantage of models based on decision trees is that decision tree models are not sensitive to the particular range or *shape* of a feature distribution. Features such as distance, household income, and household density are not naturally expressed with similar ranges of numbers and are far from uniformly distributed. For example it is common in machine learning to transform features such as household income by normalizing or taking the log of the value, or by computing a fraction into the cumulative household income distribution, to linearize these non-uniform features in order to make them more suitable for models such as linear regression, neural nets, or SVMs with linear or non-linear kernels. Decision trees, however, because they are based on threshold cuts on feature values, are not sensitive to the ranges nor shapes of feature distributions and are unaffected by monotonic feature transformations. Thus there is no need to find suitable transformations for features such as income or distance prior to training the model. This makes applying decision trees to data sets with non-uniform features easier. It also makes it easier to interpret the results of experiments that measure the importance of features, because they are not biased by the feature distribution. We explore the predictive power of various features in the next section.

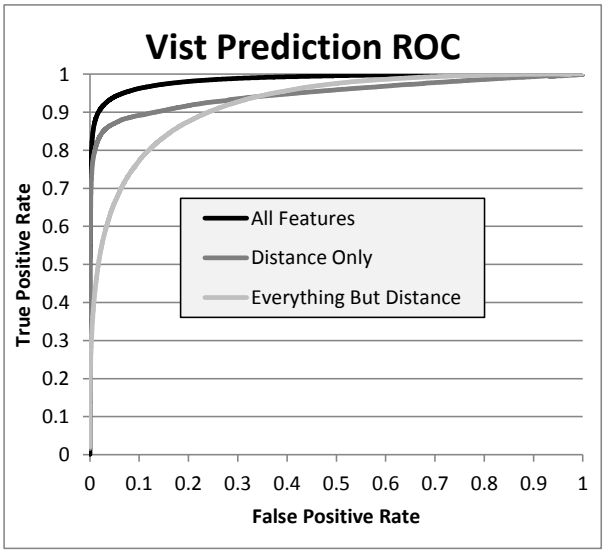

Figure 2: Receiver operating characteristic curve for visit classifier. Using all features works best, followed by distance only, and then all features but distance.

5 Results and Discussion

In this section we discuss classification accuracy and look for which features work best to find likely travel destinations.

Using two-fold cross validation, our average overall classification accuracy was 0.95. This means that our classifier correctly classified 95% of the test visits. Since our classifier produces a

probability, we can set a classification threshold to trade off false positives (classifying a visit as positive when it is not) and false negatives (classifying a visit as negative when it is not). This tradeoff is embodied in the ROC curves shown in Figure 2. The ideal operating point is the upper left, which has all true positives and no false positives. Using all 61 features comes close to this point. Using distance alone works well, too, followed by using all the features but distance. Section 5.2 examines more closely the predictive power of certain features.

We were careful not to split users' data in our cross validation runs. If we made an arbitrary 50/50 split for two-fold cross validation, it may be that the data from some users would be split between testing and training, which may give an artificial accuracy boost. Instead, we made sure that the users in the testing and training datasets were non-overlapping.

5.1 Example Maps

Figure 3 shows results of our classification algorithm. The first map shows places likely visited by someone living in Redondo Beach, California, whose location is shown as a large disk. The smaller disks show the centers of all the block groups that were classified as a destination by someone living in Redondo Beach. There is a dense cluster in the lower 2/3 of California, and another strong cluster along the coast in the northeast U.S. The second map shows likely destinations of someone living in Manhattan, NY. Here we see results clustered around the home location, and also up and down both coasts. The third map shows destinations for someone whose home is in the small town of Milburn, Nebraska (population 55). Unlike the previous two maps, many of Milburn's likely destinations are in the so-called flyover states in the middle of the country. For all three home locations, popular destinations are southern California, Florida, and the area around New York City and Washington, D.C.

The analysis above looks at the probability of all destinations given a certain home location. We can also reverse the condition and look at the probability of all home locations given a certain destination. For the given destination, this shows which locations contain people who would want to travel there. Figure 4 shows some samples. The first map suggests that people in every part of the U.S. would find it attractive to

Home in Redondo Beach, California | Home in Manhattan, New York | Home in Milburn, Nebraska (population 55)

Figure 3: Destination predictions for three different home locations. The home locations are shown as large disks, and the visited locations are shown as smaller disks, with darker ones more probable. Depending on the home location, the predicted destinations change significantly, with a cluster around the home location.

Redondo Beach has strong, wide appeal | Minneapolis, Minnesota has regional appeal | Milburn has more limited appeal

Figure 4: Classifying which home locations find a given destination attractive. The destination locations are shown as large disks, and the home locations are shown as smaller disks, with darker ones more probable.

visit Redondo Beach. In fact, the classification results give a high probability of visiting Redondo Beach from *every* block group in the U.S., with the lowest probability being 0.93. The second map shows that Minneapolis, Minnesota has regional appeal to the neighboring states. The third map in Figure 4 suggests that tiny Milburn, Nebraska appeals to mostly people from the same state. An analysis like this could be useful for targeting travel-related ads to those most likely to respond positively.

One of the reasons we chose Milburn, Nebraska as an example is that we have only four tweets from its block group, all from two users who appear to have uniformly covered most of the U.S. with geotagged tweets. There is only one user whose home was placed in Milburn's block group. While this is not nearly enough raw data to make predictions about Milburn's travel likelihoods, we can still make a reasonable prediction for this home region. Our features group Milburn with other regions of similar demographic and spatial features, meaning we can overcome the paucity of available data. This is one of the advantages of abstracting away the raw latitude/longitude locations in favor of descriptive features. The abstraction also helps us understand which features affect people's travel choices, which we explore next.

5.2 Feature Subsets

While it is gratifying to know we can classify accurately, it is more interesting to know which features lead to this level of accuracy. This is the core of our contribution, because it begins to answer, for the first time, how a person might choose their travel destinations. This is the natural next investigative step to take after much past research on the numerical statistics of human mobility.

We can easily assess the classification power of single features by rerunning our learning and test procedure on each feature individually. For instance, we can compute the classification accuracy using distance alone, median home income alone, *etc.* The results of this experiment are show in Table 2. This shows that distance is the best individual feature in terms of classification accuracy. The 19 worst features give classification accuracies of 0.5, which means they are no better than random guessing for this two-class problem. The second best feature is "Race Not Hispanic Asian Fraction (destination)", which implies that the fraction of Asians at the destination is a

Table 2: Classification accuracy of individual features. Distance is the best single feature.

Feature	Classification Accuracy
Distance	0.91
Race Not Hispanic Asian Fraction (destination)	0.67
Age 20-29 Fraction (destination)	0.64
Race Not Hispanic White Fraction (home - destination)	0.64
Race Not Hispanic White Fraction (destination)	0.63
Age 30-39 Fraction (destination)	0.62
Age 10-19 Fraction (destination)	0.62
Household Density (destination)	0.62
Race Hispanic Fraction (destination)	0.62
Race Not Hispanic Two or More Fraction (destination)	0.61
Race Not Hispanic Other Fraction (destination)	0.61
Race Not Hispanic Asian Fraction (home - destination)	0.61
Age Under 10 Fraction (destination)	0.60
Age Over 69 Fraction (destination)	0.60
Age 60-69 Fraction (destination)	0.60
Race Hispanic Fraction (home - destination)	0.60
Race Not Hispanic Black Fraction (destination)	0.60
Age 50-59 Fraction (destination)	0.60
Race Not Hispanic Black Fraction (home - destination)	0.59
Race Not Hispanic Two or More Fraction (home - destination)	0.59
Age 20-29 Fraction (home - destination)	0.59
Race Not Hispanic Islander Fraction (destination)	0.58
Age 60-69 Fraction (home - destination)	0.57
Household Density (home - destination)	0.57
Age Over 69 Fraction (home - destination)	0.57
Race Not Hispanic Other Fraction (home - destination)	0.57
Mean Household Size (destination)	0.57
Median Income (destination)	0.57
Land Area Fraction (home - destination)	0.57
Age 10-19 Fraction (home - destination)	0.57
Age 40-49 Fraction (destination)	0.57

Feature	Classification Accuracy
Age 50-59 Fraction (home - destination)	0.57
Land Area Fraction (destination)	0.56
Age Under 10 Fraction (home - destination)	0.56
Age 30-39 Fraction (home - destination)	0.56
Race Not Hispanic Islander Fraction (home - destination)	0.55
Race Not Hispanic Indian Fraction (home - destination)	0.54
Age 40-49 Fraction (home - destination)	0.53
Median Income (home - destination)	0.53
Mean Household Size (home - destination)	0.53
Race Not Hispanic Indian Fraction (destination)	0.52
Race Not Hispanic Other Fraction (home)	0.50
Race Not Hispanic Islander Fraction (home)	0.50
Race Not Hispanic Indian Fraction (home)	0.50
Race Not Hispanic Two or More Fraction (home)	0.50
Race Not Hispanic Asian Fraction (home)	0.50
Age 40-49 Fraction (home)	0.50
Land Area Fraction (home)	0.50
Age 50-59 Fraction (home)	0.50
Age 60-69 Fraction (home)	0.50
Age Over 69 Fraction (home)	0.50
Race Hispanic Fraction (home)	0.50
Age 20-29 Fraction (home)	0.50
Age 30-39 Fraction (home)	0.50
Race Not Hispanic Black Fraction (home)	0.50
Age Under 10 Fraction (home)	0.50
Race Not Hispanic White Fraction (home)	0.50
Age 10-19 Fraction (home)	0.50
Mean Household Size (home)	0.50
Household Density (home)	0.50
Median Income (home)	0.50

relatively important feature, although it only achieves a classification accuracy of 0.67 when used alone.

If a feature works well by itself, then it will continue to work well in combination with others. The opposite is not true: if a feature does not work well by itself, it can still be effective in combination with others. Thus we next examine groups of features.

This analysis of individual features suggests that some groups of features may be more powerful than others. For instance, in Table 2, we see that destination features are generally ranked above home features. Also, after the distance feature, we see a mix of race and age features until "Mean Household Size (destination)" at number 27.

We can assess groups of features in the same way, by training and testing our classifier on certain subsets of features. Table 3 shows the results of this experiment for some manually chosen feature subsets. The constituent features in most of the groups are obvious by the group names. Note that the Race feature group contains all the race fraction features, and the Age feature group contains all the age range features. Only "All Features" and "Distance" have the distance feature. Other groups include those features computed from the home and destination block groups as well as the differences between features in these block groups.

We see that using all features gives an accuracy of 0.95, followed closely by distance alone at 0.91. Combined with our analysis of individual features above, this shows that distance is the dominant feature determining travel. Equally interesting, however, is the relatively high accuracy of the all-but-distance group at 0.84. This shows that demographic and spatial features are important considerations in picking a travel destination. While the distance feature alone works well, it is helped by non-distance features, boosting accuracy from 0.91 to 0.95. We also see that race is an important consideration, with a classification accuracy of 0.81. While this seems a somewhat uncomfortable conclusion, it may well be that race is not an explicit consideration when choosing a destination, but serves as a proxy for visiting relatives or

Table 3: Classification accuracy of selected groups of features. Omitting distance still gives good accuracy. Race is somewhat high, while features around the home are not important when used alone in a group.

Feature Group	Classification Accuracy
All Features	0.95
Distance	0.91
All But Distance	0.84
Race	0.81
Destination (All Features)	0.79
Home - Destination (All Features)	0.79
Race at Destination	0.74
Age	0.73
Race (Home - Destination)	0.73
Age at Destination	0.71
Age (Home - Destination)	0.67
Household Density	0.65
Income	0.60
Land Fraction	0.59
Household Size	0.59
Race at Home	0.49
Age at Home	0.49
Home (All Features)	0.49

cultural events. We note that "Race at Destination" ranks significantly higher than "Race at Home". Home features generally rank lower than those involving the destination, with "Home (All Features)" (all features around home) doing no better than random guessing. Apparently, the destination is a more important consideration than the person's home when choosing where to go.

While the analysis above shows which features are important in predicting destinations, it does not show *how* the features affect a destination's likelihood. For instance, we have shown that distance is an important predictive feature, but we don't yet know if smaller distances increase or decrease the likelihood of a visit. For individual features, we can answer this question by looking at the distributions of the feature's values for both positive and negative examples. As we expect, for the distance feature, the mean distance of the positive examples is less than the mean distance for the negative examples, and this difference in means is statistically significant. This means that nearby destinations are more attractive than distant ones. We did this test on all 61 individual features, and the results are shown in Table 4. For each feature, we computed the means of the positive and negative examples. We used a two-sample t-test to accept or reject the hypothesis that the means were equal. In Table 4, an arrow indicates that the difference in means for a feature was statistically significant at the α=0.05 level. If the arrow points up, then the mean of the positives was greater than the mean of the negatives. For instance, the up arrow for "Median income" at the destination means that the median income at positive sample destinations was higher than that at negative samples. Apparently higher median income makes a destination more attractive. We can also look at the difference (home – visit) of "Median Income", which has a down arrow. This feature looks at the difference in median income between the home and destination, and the down arrow implies that a smaller difference makes the visit more likely. An interesting pattern in Table 4 is that all the purely "home" features are not statistically significant, implying that they have little effect on the choice of a destination. This supports a similar conclusion we made after looking at the predictive power of the whole group of purely "home" features in Table 3. We also note that a higher proportion of people in the age range 20-39 increases the attractiveness of a destina-

Table 4: Individual features' effects on visit likelihood. An up arrow indicates that the mean of the feature is higher for visits than it is for non-visits. A down arrow indicates that the mean feature value is lower for visits. An × indicates that the difference in means was not statistically significant. None of the purely "home" features showed a significant difference in the means.

Feature Group	Feature Name	Description	Home	Dest.	Diff.
Distance	Distance	Distance between home and destination	N/A	N/A	↓
Income	Median income	Median income in U.S. dollars	×	↑	↓
Race	Non-Hispanic White	Fraction non-Hispanic whites	×	↓	↑
	Non-Hispanic Black	Fraction non-Hispanic blacks	×	↑	↓
	Non-Hispanic Indian	Fraction non-Hispanic Indians	×	↓	↑
	Non-Hispanic Asian	Fraction non-Hispanic Asians	×	↑	↓
	Non-Hispanic Islander	Fraction non-Hispanic islanders	×	↑	↓
	Non-Hispanic Other	Fraction non-Hispanic other than above	×	↑	↓
	Non-Hispanic Two	Fraction non-Hispanic two or more races	×	↑	↓
	Hispanic	Fraction Hispanic	×	↑	↓
Age	Under 10	Fraction under 10 years old	×	↓	↑
	10-19	Fraction 10-19 years old	×	↓	↑
	20-29	Fraction 20-29 years old	×	↑	↓
	30-39	Fraction 30-39 years old	×	↑	↓
	40-49	Fraction 40-49 years old	×	↓	↑
	50-59	Fraction 50-59 years old	×	↓	↑
	60-69	Fraction 60-69 years old	×	↓	↑
	Over 69	Fraction over 69 years old	×	↓	↑
Household Size	Mean household size	Mean number of persons in household	×	↑	↓
Household Density	Houshold Density	Households per unit land area	×	↑	↓
Land Fraction	Land Fraction	Fraction of land area (as opposed to water)	×	↓	↑

tion, while people in other age ranges tends to decrease its attractiveness. To the best of our knowledge, this is the first systematic examination of how various features affect the likelihood of a visit.

6 Conclusions

We have shown that we can accurately predict which places a person will visit based on the travel distance, demographics, and spatial features of the person's home and destination. Our experiments were based on over 3.2 million visits made by over 200,000 Twitter users. These experiments are some of the first to attempt to understand how people pick destinations, going beyond just statistics about their travel. We found that we could predict destinations down to the Census block level with 95% accuracy, or 84% accuracy if we ignore travel distance. We found that distance is the dominant feature in predicting where someone will go, but that non-distance features still perform well, and they help in combination with distance. In examining groups of features, we found that features around the home are in general less powerful predictors than features around the destination. We also showed how the relative values of features affect the likelihood of a visit.

We see our work as one of the first steps toward automatically understanding how people pick travel destinations. Future work should examine the role of other features, like businesses, attractions, weather, and seasons, which we feel will be important.

7 Acknowledgments

Thank you to the TLC team at Microsoft Research for providing their machine learning package and to Ed Katibah of SQL Server for spatial data.

References

1. Eagle, N. and A. Pentland, *Reality mining: sensing complex social systems.* Personal and Ubiquitous Computing, 2006. **10**(4): p. 255-268.

2. Bayir, M.A., M. Demirbas, and N. Eagle, *Discovering SpatioTemporal Mobility Profiles of Cellphone Users*, in *10th IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM 2009)*. 2009: Kos, Greece. p. 1-9.
3. González, M.C., C.A. Hidalgo, and A.-L. Barabási, *Understanding individual human mobility patterns.* Nature, 2008(453): p. 779-782.
4. Brockmann, D., L. Hufnagel, and T. Geisel, *The scaling laws of human travel.* Nature, 2006(439): p. 462-465.
5. Isaacman, S., et al., *Human Mobility Modeling at Metropolitan Scales*, in *Tenth Annual International Conference on Mobile Systems, Applications, and Services (MobiSys 2012)*. 2012: Low Wood Bay, Lake District, UK.
6. Cheng, Z., et al., *Exploring Millions of Footprints in Location Sharing Services*, in *5th International AAAI Conference on Weblogs and Social Media (ICWSM)*. 2011: Barcelona, Spain.
7. Krumm, J., *Inference Attacks on Location Tracks*, in *Fifth International Conference on Pervasive Computing (Pervasive 2007)*. 2007, Springer: Toronto, Ontario Canada. p. 127-143.
8. Friedman, J.H., *Greedy Function Approximation: A Gradient Boosting Machine.* The Annals of Statistics, 2001. **29**(5): p. 1189-1232.
9. Caruana, R. and A. Niculescu-Mizil, *An Empirical Comparison of Supervised Learning Algorithms*, in *23rd International Conference on Machine Learning (ICML 2006)*. 2006: Pittsburgh, PA. p. 161-168.