Bayesian estimates for trajectory matching

Dániel Kondor, Senseable City Lab, MIT

February 27, 2017

In this report I summarize some basic findings about estimating the probability of two spatial trajectories from two different datasets belonging to the same person. These complement the analysis in the upcoming manuscript [1]. The main application is for the same two datasets described there. I use estimates based on the Bayes theorem:

$$P(A|B) = \frac{P(A)}{P(B)} P(B|A) \tag{1}$$

The basic ideas are the same as in [2], while the applicability of assumptions about independence among different events will not be applicable in our case.

1 Basic estimate

In our case we would like to estimate $P(\mathbf{x}_i \text{matches} \mathbf{y}_j|\text{data})$, i.e. the probability that trajectory $\mathbf{x}_i$ from dataset X belongs to the same real person as trajectory $\mathbf{y}_j$ from dataset Y, given all the records we have in the two datasets, which is at this point represented by the abstract "data" requirement in the conditional. We apply Bayes theorem to this probability, gaining the following form:

$$P(\mathbf{x}_i \text{ matches } \mathbf{y}_j|\text{data}) = \frac{P(\mathbf{x}_i \text{ matches } \mathbf{y}_j)}{P(\text{data})} P(\text{data}|\mathbf{x}_i \text{ matches } \mathbf{y}_j) \tag{2}$$

Here, we can estimate the three terms on the right hand side more easily. We can perform an estimation of $P(\text{data}|\mathbf{x}_i \text{matches} \mathbf{y}_j)$ based on the datasets. The term $P(\mathbf{x}_i \text{ matches } \mathbf{y}_j)$ we identify as *a priori* probabilities, i.e. the probability that any two trajectories match in the two datasets without any further knowledge. We can simply define it to be $P(\mathbf{x}_i \text{ matches } \mathbf{y}_j) = 1/N$, where N is the number of individuals in either dataset X or Y or more generally, the population present in the area where the trajectories were recorded (this allows the datasets to be incomplete, e.g. when they come from a service provider with a $< 100\%$ market share in the area). The denominator, $P(\text{data})$ denotes the probability of arriving at the datasets at hand; we can eliminate this by considering also the probability of $\mathbf{x}_i$ and $\mathbf{y}_j$ not being a match and exploiting that the two probabilities has to sum up to 1 (i.e. two trajectories either belong to the same person or not):

$$P(\mathbf{x}_i \text{ does not match } \mathbf{y}_j|\text{data}) = \frac{P(\mathbf{x}_i \text{ does not match } \mathbf{y}_j)}{P(\text{data})} P(\text{data}|\mathbf{x}_i \text{ does not match } \mathbf{y}_j) \tag{3}$$

and

$$P(\mathbf{x}_i \text{ matches } \mathbf{y}_j|\text{data}) + P(\mathbf{x}_i \text{ does not match } \mathbf{y}_j|\text{data}) = 1 \tag{4}$$

Replacing both terms with the form obtained from the Bayes theorem, and using the shorthand notations for the conditional probabilities on the right hand sides: $P_T \equiv P(\text{data}|\mathbf{x}_i \text{ matches } \mathbf{y}_j)$ and $P_S \equiv P(\text{data}|\mathbf{x}_i \text{ does not match } \mathbf{y}_j)$, we get the following expression:

$$P(\text{data}) = \frac{P_T}{N} + \frac{N-1}{N} P_S \tag{5}$$

where we also used the a priori probabilities $P(\mathbf{x}_i \text{ matches } \mathbf{y}_j) = 1/N$ and $P(\mathbf{x}_i \text{ does not match } \mathbf{y}_j) = (N-1)/N$. This results in the following form for the original probability of match to be estimated:

$$P(\mathbf{x}_i \text{ matches } \mathbf{y}_j|\text{data}) = \frac{P_T}{N} \frac{1}{\frac{P_T}{N} + \frac{N-1}{N} P_S} = \frac{1}{1 + (N-1)\frac{P_S}{P_T}} = \frac{P_T}{P_T + (N-1) P_S} \tag{6}$$

In the following, the main challenge is estimating P_T and P_S realistically from the datasets.

2 Empirical match distributions

In this section, I base the estimation of P_T and P_S on the empirical distribution of getting a certain m number of matches between a pair of users as estimated from the datasets (described in [1]). The main approximation I use is that the difference between P_T and P_S is only whether the two trajectories ($\mathbf{x}_i$ and $\mathbf{y}_j$) actually belong to the same person or not. Note that P_T and P_S actually includes the probability of getting all the records in both the datasets. In the following, I use m to denote the number of matching records between the two trajectories (the criterion for matching is that the point are sufficiently close in both time and space). Note that m is determined from the data, i.e. it is a parameter for estimating the probabilities, being a part of the "data" abstraction. I assume that the trajectories have no inconsistent records (i.e. points which are close in time but separated in space); if there are, then the probability of them being a true match can be trivially determined to be zero. In this section, I then proceed with writing these probabilities in product forms:

$$P_T = P_{ij}(m|\text{match})P(\text{data} \setminus (\mathbf{x}_i, \mathbf{y}_j)|\text{match}) \tag{7}$$
$$P_S = P_{ij}(m|\text{not match})P(\text{data} \setminus (\mathbf{x}_i, \mathbf{y}_j)|\text{not match}) \tag{8}$$

In these cases, P_{ij} is the probability of finding m (spatially consistent) matches between the two trajectories ($\mathbf{x}_i$ and $\mathbf{y}_j$), given that these belong to the same person ("match") or not ("not match"). The second terms in the products are the probabilities of getting the rest of the records in these cases. To be able estimate these, I make the following asusmptions:

1. $P(\text{data}\setminus(\mathbf{x}_i, \mathbf{y}_j)|\text{match}) = P(\text{data}\setminus(\mathbf{x}_i, \mathbf{y}_j)|\text{not match})$, i.e. the distribution describing the rest of the data does not depend on whether the two trajectories are a real match or not. This allows us to completely disregard this term, as it can be eliminated from equation 6. Note that this is only an approximation; if the two trajectories are not a real match, both can have their true match in the other dataset, resulting in the trajectories having a higher number of matches with some other trajectory.

2. The P_{ij} distributions do not depend on the specific pair of trajectories selected, but only on the number of records in each of them. This is essentially the same assumption as the ones presented in Section 5.1. in the main manuscript [1]. This assumption is necessary to estimate the probabilities from the empirical distributions in the data.

3. The $P_{ij}(m|\text{match})$ distribution essentially gives the probability of having m *temporal* matches between the matching trajectories of a person (since in this case it is assumed that all records are spatially consistent then). I then assume that this can be estimated with the empirical distribution of having m temporal matches between the records of *any* pair of trajectories from the two datasets, even if those are not spatially consistent.

Based on these asusmptions, and omitting the second terms (which cancel out), the probabilities can be written in the following form, using the notations from Section 5.1. in [1]:

$$P_T = P_t(m|r_i, r_j) \tag{9}$$
$$P_S = P_s(m|r_i, r_j) \tag{10}$$

Here, P_t and P_s are empirical distributions between groups of trajectories with r_i and r_j records as estimated from the data: $P_t(m|r_i, r_j)$ is the distribution of temporal matches between this group of trajectories, while $P_s(m|r_i, r_j)$ is the distribution of spatially consistent matches in this group. Using these estimates, I display the obtained probabilities for different user groups and number of matches in Fig. 1. We see that the probabilities

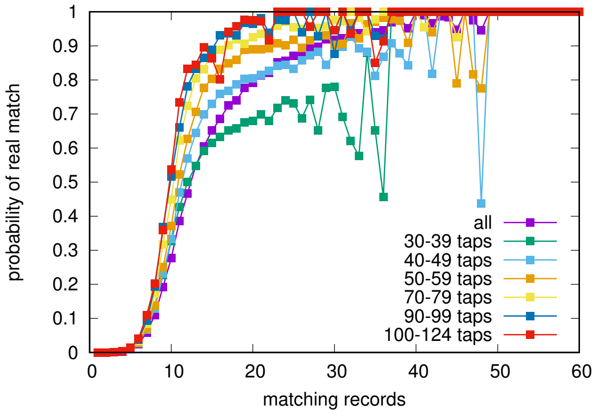

Figure 1: **Average estimate probability of a true match between two trajectories as a function of the number of matching records.** Results are displayed as an average for all user records in the CDR dataset, while the separate lines show results for different groups of transportation users (the purple line is an average for any pair of trajectories in the two datasets).

display a rapid transition around 10-15 matching records (which correspong to about 50% probability), while to reach high confidence, a significantly high number of matching records are required, i.e. around or above 30. Looking at Eq. 6, we see that the main requirement of having a match with high confidence is that $P_T \gtrsim NP_S$, i.e. the probability of the real match of the two trajectories has to be larger than the probability of a false positive match with *any* of the N trajectories in the dataset (note that an exact formula for having at least one false positive matching trajectory with any out of N would be $1-(1-P_S)^N \approx NP_S$ for $NP_S \ll 1$).

3 Independent individual points

The main drawback of the previous approach is that it requires a good approximation of the P_T and P_S distributions based on empirical data. Especially for P_S and a large number of matches, this becomes difficult as there are only a few such matching trajectories in the real dataset. This estimation seems much more problematic than estimating the p_x average expected success ratios as done in Ref. [1]. Also, the estimation relies on average probabilities and thus cannot take into account individual differences in trajectories, e.g. the fact that a matching pair of records in a less crowded area could be given more weight. For these reasons, it would be beneficial to have an estimate where probabilities are assigned to each record separately and combined together. In this section, I present the simplest such possible scenario, essentially the adaptation of the formulas in Ref. [2], and then argue that the results are probably not correct as matching probabilities for separate records are not independent (i.e. since people do not move uniformly randomly, but follow a given base routine).

In the following, I use the shorthand notation $\mathbf{x}_i \sim \mathbf{y}_j$ to denote that the two trajectories belong to the same person and $\mathbf{x}_i \nsim \mathbf{y}_j$ to denote that they do not. Furthermore, I assume that the $\mathbf{x}_i$ trajectory has r_i points, and denote by z_k the total set of matching points around point x_{ik} in trajectory $\mathbf{x}_i$ ($k \in 1, 2, \ldots r_i$). In this scenario, I then assume that the P_T and P_S probabilities can be written as the products of individual probabilities corresponding to each point in the trajectories and then apply the Bayes theorem to each of these separately, essentially using the same approach as described in the Appendix of Ref. [2]:

$$P_T = \prod_{k=1}^{r_i} P(z_k|\mathbf{x}_i \sim \mathbf{y}_j) = \prod_{k=1}^{r_i} \frac{P(z_k)}{P(\mathbf{x}_i \sim \mathbf{y}_j)} P(\mathbf{x}_i \sim \mathbf{y}_j|z_k) \tag{11}$$

$$P_S = \prod_{k=1}^{r_i} P(z_k|\mathbf{x}_i \nsim \mathbf{y}_j) = \prod_{k=1}^{r_i} \frac{P(z_k)}{P(\mathbf{x}_i \nsim \mathbf{y}_j)} P(\mathbf{x}_i \nsim \mathbf{y}_j|z_k) \tag{12}$$

$$P(\mathbf{x}_i \sim \mathbf{y}_j|\text{data}) = \frac{P(\mathbf{x}_i \sim \mathbf{y}_j)}{P(\text{data})} P_T = \frac{\prod P(z_k)}{P(\text{data})} \frac{\prod P(\mathbf{x}_i \sim \mathbf{y}_j|z_k)}{P(\mathbf{x}_i \sim \mathbf{y}_j)^{r_i - 1}} \tag{13}$$

$$P(\mathbf{x}_i \nsim \mathbf{y}_j|\text{data}) = \frac{P(\mathbf{x}_i \nsim \mathbf{y}_j)}{P(\text{data})} P_S = \frac{\prod P(z_k)}{P(\text{data})} \frac{\prod P(\mathbf{x}_i \nsim \mathbf{y}_j|z_k)}{P(\mathbf{x}_i \nsim \mathbf{y}_j)^{r_i - 1}} \tag{14}$$

Again, the first factor can be eliminated using the relation that $P(\mathbf{x}_i \sim \mathbf{y}_j|\text{data}) + P(\mathbf{x}_i \nsim \mathbf{y}_j|\text{data}) = 1$:

$$\frac{P(\text{data})}{\prod P(z_k)} = \frac{\prod P(\mathbf{x}_i \sim \mathbf{y}_j|z_k)}{P(\mathbf{x}_i \sim \mathbf{y}_j)^{r_i - 1}} + \frac{\prod P(\mathbf{x}_i \nsim \mathbf{y}_j|z_k)}{P(\mathbf{x}_i \nsim \mathbf{y}_j)^{r_i - 1}} \tag{15}$$

Using this, and performing some simplifications, and also using that $P(\mathbf{x}_i \sim \mathbf{y}_j|z_k) + P(\mathbf{x}_i \nsim \mathbf{y}_j|z_k) = 1$ and $P(\mathbf{x}_i \sim \mathbf{y}_j) + P(\mathbf{x}_i \nsim \mathbf{y}_j) = 1$, we have the following result:

$$P(\mathbf{x}_i \sim \mathbf{y}_j|\text{data}) = \left(1 + \left(\frac{P(\mathbf{x}_i \sim \mathbf{y}_j)}{1 - P(\mathbf{x}_i \sim \mathbf{y}_j)}\right)^{r_i - 1} \prod_{k=1}^{r_i} \frac{1 - P(\mathbf{x}_i \sim \mathbf{y}_j|z_k)}{P(\mathbf{x}_i \sim \mathbf{y}_j|z_k)}\right)^{-1} \tag{16}$$

Similarly to the previous case, we set $P(\mathbf{x}_i \sim \mathbf{y}_j) = 1/N$, while the probability of having a match can be estimated very similarly to the scheme in Ref. [2]:

$$P(\mathbf{x}_i \sim \mathbf{y}_j|z_k) = \begin{cases} \frac{1}{N} & \text{if no matching point for in } \mathbf{y}_j \text{ for } x_{ik} \\ \frac{1}{p_k N} & \text{if } \exists y_{jl} \text{ spatially consistent match } x_{ik} - y_{jl} \\ 0 & \text{if } \exists y_{jl} \text{ impossible match } x_{ik} - y_{jl} \end{cases} \tag{17}$$

where p_k is the ratio of spatially consistent matches to all temporal matches near point x_{ik}. The rationale for this choice is illustrated in Fig. 2 and is very similar to the way probabilities were estimated in Ref. [2] (see Fig. 4 there). For this discussion, we use the same definitions of matching as in the main manuscript (Ref. [1], see Section 4.1.): we define two records (from the two different datasets) to be a *temporal match* if they are close to each other in time (closer than a threshold τ, defined as five minutes when traveling by transit and ten minutes when walking). Among temporal matches, we refer to records as *spatial matches* if their distance is less than a given d threshold (chosen as 500 m for walking and 2 km for using transit). We refer to records outside this threshold a *impossible matches.* In Eq. 17, we are calculating the probability of trajectories $\mathbf{x}_i$ and $\mathbf{y}_j$ belonging to a same person, given only the knowledge about z_k, i.e. the points in the neighborhood of $x_{ik} \in \mathbf{x}_i$. If trajectory $\mathbf{y}_j$ has no points as a temporal match there, then we learn no extra knowledge, resulting in using the a priory probability of $1/N$. In the case there is a point in the set of impossible matches, we can rule out them belonging to the same person, hence the probability becomes zero (and note that this will imply $P(\mathbf{x}_i \sim \mathbf{y}_j|\text{data}) = 0$ as well). On the other hand, if the two trajectories have a spatial match, this provides further evidence for these to belong to the same person. The factor of p_k then essentially represents a "redistribution" of probabilities: assuming that the total number of trajectories (in Y) is N, and we have N_k^T with temporal matches, N_k^I of these impossible, and N_k^S spatial matches ($N_k^I + N_k^S = N_k^T$), then the total probability among N_k^T trajectories is redistributed among only N_k^S, resulting in the $p_k \equiv N_k^S / N_k^T$ factor in comparison to the a priory probabilities.

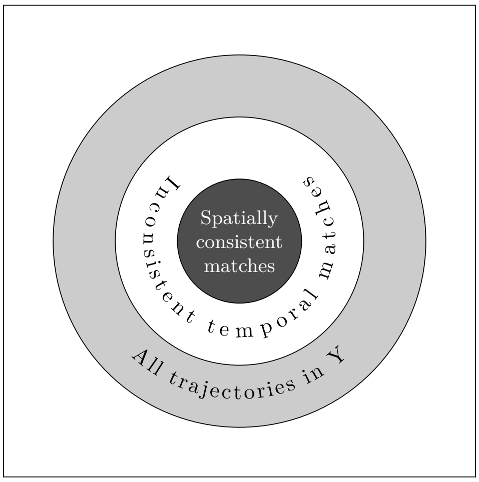

Figure 2: **Relation of the different types of trajectories.** The sets correspond to trajectories from Y with relation to one point x_{ik} from a trajectory in X. The darker the subset is, the higher the probability of a match. See Fig. 4 and the related discussion in Ref. [2] for reference.

Using the probabilities from Eq. 17, and noticing that we can eliminate the records from $\mathbf{x}_i$ which have no temporal matches in $\mathbf{y}_j$, we arrive at the following formula:

$$P(\mathbf{x}_i \sim \mathbf{y}_j|\text{data}) = \frac{1}{1 + \left(\frac{1}{N-1}\right)^{m-1} \prod_{k=1}^{m}(p_k N - 1)} \tag{18}$$

where trajectories $\mathbf{x}_i$ and $\mathbf{y}_j$ have m spatially consistent matches, and $k = 1, 2, \ldots m$ now indexes the m records in $\mathbf{x}_i$ which have these matching points in $\mathbf{y}_j$. What remains in this case is that for each match, we have to characterize the neighborhood of each matching point. In the following, I show a simplified approximation where I assume that all p_m ratios can be approximated from above. Clearly, if we assume that $p_m < p \forall m$, we have the following result for the estimate of the previous equation:

$$P(\mathbf{x}_i \sim \mathbf{y}_j|\text{data}) > \frac{1}{1 + (N-1)\left(\frac{pN-1}{N-1}\right)^{m}} \approx \frac{1}{1 + Np^m} \tag{19}$$

where the last approximation is simply based on that $N \gg 1$ and $p \ggg 1/N$. Note that for a given m, the probability estimate will be increasing as p decreases. In Fig. 3, I display the result in Eq. 19 along with the previous estimate as well. We see that estimates calculated from Eq. 19 give much higher values for match probabilities than the previous approach which based estimates on the empirical distribution of matches among trajectories in the data. I believe that the previouy methodology is more correct and the results based on Eq. 19 actually overestimate the real probabilities. The main approximation which is questionable is writing P_T and P_S as a product of independent probabilities for each record in $\mathbf{x}_j$. While in the case of P_T, this could be justified by the assumptions that generating a record in the two datasets is independent, but in the case of P_S, the probability of record pairs being spatially consistent is not independent: we can argue that this depends on a latent parameter whether the people generating the two trajectories have similar mobility habits (i.e. live and work relatively close to each other), as human mobility in general shows high regularity.

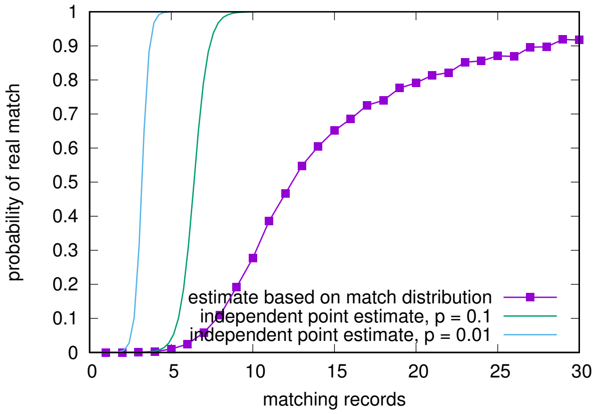

Figure 3: **Average estimate probability of a ture match between two trajectories as a function of the number of matching records.** The two lines are based on Eq. 19, with p values of 0.1 and 0.01 (the smaller value results in higher probabilities. We expect real p_m values to be even smaller, but these values already result in much higher match probabilities than the previous estimation method. The previous estimate (the same data as in Fig. 1) is shown for comparison.

References

[1] Kondor D, Hashemian B, de Montjoye YA, Ratti C (2017). Towards matching user mobility traces in large-scale datasets. *manuscript in preparation* `https://www.dropbox.com/s/gm51conqrvalv4z/bigdatapaper_merged.pdf?dl=0`

[2] Juhász PL, Stéger J, Kondor D, Vattay G (2016). A Bayesian Approach to Identify Bitcoin Users. `arXiv:1612.06747`. Retrieved from `http://arxiv.org/abs/1612.06747`