Minimum Length of Sensor Data Collection for Robust Mobility Estimation

Zhihang Dong, Yen-Chi Chen, Adrian Dobra

Department of Statistics
University of Washington, Seattle, USA
Joint Statistical Meeting 2018
Vancouver, B.C., Canada
July 30th, 2018

CENTER for STATISTICS and the SOCIAL SCIENCES

Authors

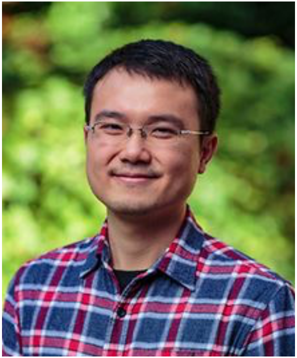


FIGURE – Yen-Chi

Authors

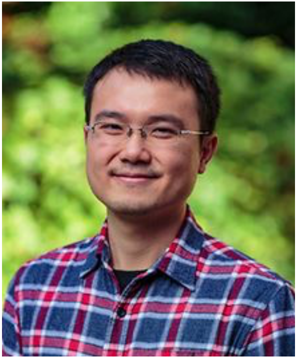


FIGURE – Yen-Chi

FIGURE – Adrian

Authors

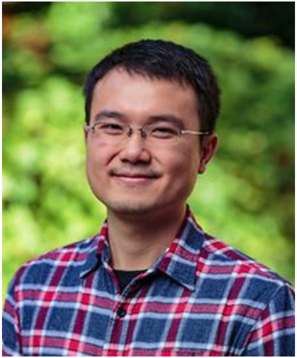


FIGURE – Yen-Chi

FIGURE – Adrian

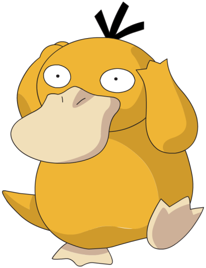


FIGURE – Zhihang (Presenter)

A Fundamental Question

The development of pervasive computing and wearable sensor technology has brought up an exponential growth of data of human activities. With great data comes with great responsibilities... How to handle these data ?

FIGURE – Wearable Sensor Devices (Fitbit)

Current Research

There are many topics studied... Here is a group of topics funded by NIH using censor data...

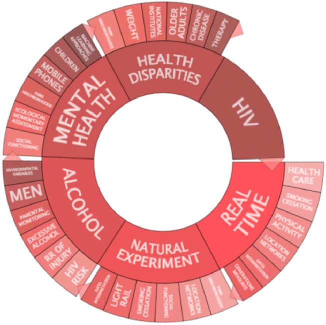


FIGURE – Topics using Sensor Data

A Fundamental Question (2)

Most of the current research has addressed one of the five stages in human activity data :

- Stage 1 : *Sourcing and System* : Pervasive Computing (Althoff, 2017)

A Fundamental Question (2)

Most of the current research has addressed one of the five stages in human activity data :

- Stage 1 : *Sourcing and System* : Pervasive Computing (Althoff, 2017)
- Stage 2 : *Ethics* : Privacy, Data Ownership and Protocols... (Harari et al. 2016)

A Fundamental Question (2)

Most of the current research has addressed one of the five stages in human activity data :

- Stage 1 : *Sourcing and System* : Pervasive Computing (Althoff, 2017)
- Stage 2 : *Ethics* : Privacy, Data Ownership and Protocols... (Harari et al. 2016)
- Stage 3 : *Validation* : Variance, Bias, Spatio-Temporal Complexity and beyond... (Kwan 2009, 2012 and Kwan et al. 2003)

A Fundamental Question (2)

Most of the current research has addressed one of the five stages in human activity data :

- Stage 1 : *Sourcing and System* : Pervasive Computing (Althoff, 2017)
- Stage 2 : *Ethics* : Privacy, Data Ownership and Protocols... (Harari et al. 2016)
- Stage 3 : *Validation* : Variance, Bias, Spatio-Temporal Complexity and beyond... (Kwan 2009, 2012 and Kwan et al. 2003)
- Stage 4 : *Methodology* : A Thousand Methods (Inference, ML, DL...)

A Fundamental Question (2)

Most of the current research has addressed one of the five stages in human activity data :

- Stage 1 : *Sourcing and System* : Pervasive Computing (Althoff, 2017)
- Stage 2 : *Ethics* : Privacy, Data Ownership and Protocols... (Harari et al. 2016)
- Stage 3 : *Validation* : Variance, Bias, Spatio-Temporal Complexity and beyond... (Kwan 2009, 2012 and Kwan et al. 2003)
- Stage 4 : *Methodology* : A Thousand Methods (Inference, ML, DL...)
- Stage 5 : *Applications* : Population Health (Zenk et al. 2011, 2012), Public Safety (Graif et al. 2014), Intervention (Free et al. 2013) and Epidemiology (Wu et al. 2010)...

A Fundamental Question (3)

We are concerning Stage 3 with an important question in mind :
What is the minimum amount of time required in order to capture a (moderately) complete picture of human activity ?

- Stage 1 : *Sourcing and System* : Pervasive Computing
- Stage 2 : *Ethics* : Privacy, Data Ownership and Protocols...
- **Stage 3 : *Validation* : Variance, Bias, Spatio-Temporal Complexity...**
- Stage 4 : *Methodology* : A Thousand Methods (Inference, ML, DL...)
- Stage 5 : *Applications* : Population Health , Public Safety, Intervention and Epidemiology...

Rationale

There are several reasons why we should study this problem :

- Data may be affordable, but not free !

Rationale

There are several reasons why we should study this problem :

- Data may be affordable, but not free !
- Computing burden increased "skyrocketingly" when unnecessarily long time series are considered

Rationale

There are several reasons why we should study this problem :

- Data may be affordable, but not free !
- Computing burden increased "skyrocketingly" when unnecessarily long time series are considered
- Current health research (and many else) uses an unjustified data collection convention (usually 7 days)[see the figure coming up...]

Rationale

There are several reasons why we should study this problem :

- Data may be affordable, but not free !
- Computing burden increased "skyrocketingly" when unnecessarily long time series are considered
- Current health research (and many else) uses an unjustified data collection convention (usually 7 days)[see the figure coming up...]
- There is an absence of a metric to evaluate the coverage (Why do we need a **metric** ?)

Current Practices of Sensor Data Collection

Below is a sample of NIH/non-NIH funded projects with respect to their data collection length, sample size and publication/project funding year. Projects recruit more participants, collect longer data, and are more heavily funded.

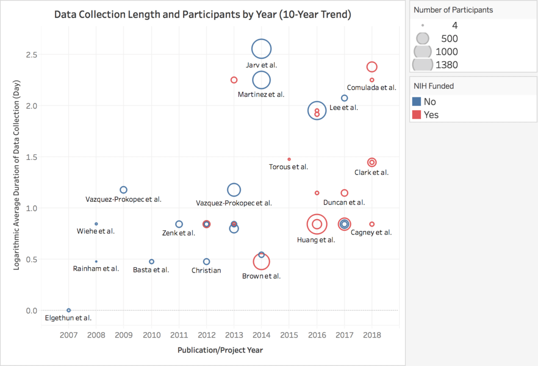


FIGURE – Current Practices of Sensor Data Collection

Notes : $e^{1.5} \approx 4.4$ days, $e^2 \approx 7$ days (1 week), $e^{2.5} \approx 12$ days(roughly two weeks)

Key Aspects to Address

To investigate a complete picture of human activity using censor data, there are some key aspects to consider :

Mobility

The variance, consistency and range of the individual travel behavior in a 'normal' day (not the day you took vacation flight to Sri Lanka).

Key Aspects to Address

To investigate a complete picture of human activity using censor data, there are some key aspects to consider :

Mobility

The variance, consistency and range of the individual travel behavior in a 'normal' day (not the day you took vacation flight to Sri Lanka).

Activity Space

The coverage of activity space (amount to the information sample space) we have using this length of data (are they important ? see Block 3).

Key Aspects to Address

To investigate a complete picture of individual activities using censor data, there are some key aspects to consider :

Mobility

The variance, consistency and range of the individual travel behavior in a 'normal' day (not the day you took vacation flight to Sri Lanka).

Activity Space

The coverage of activity space (amount to the information sample space) given the available spatial data points (are they important ? see Block 3).

Exposure

The spatio-temporal information of these data points : the length of stay, the contingency of "hotspots", etc...

Data

We will be able to address all of them using our data.
The data come from the MDC, a big data Challenge based on Lausanne, Switzerland :

- People : 185 participants participated the study with the length up to two years. Records sent to server every few seconds (or more) from their mobile devices.

Data

We will be able to address all of them using our data.
The data come from the MDC, a big data Challenge based on Lausanne, Switzerland :

- People : 185 participants participated the study with the length up to two years. Records sent to server every few seconds (or more) from their mobile devices.
- Profile : 62% male, 38% female ; a concentration of young-age population

Data

We will be able to address all of them using our data.
The data come from the MDC, a big data Challenge based on Lausanne, Switzerland :

- People : 185 participants participated the study with the length up to two years. Records sent to server every few seconds (or more) from their mobile devices.
- Profile : 62% male, 38% female ; a concentration of young-age population
- Data : Demographic (very limited) ; Phone GPS data, Wi-Fi Scans, accelerometer (sum up to about 50 million unique location points).

Data

We will be able to address all of them using our data.
The data come from the MDC, a big data Challenge based on Lausanne, Switzerland :

- People : 185 participants participated the study with the length up to two years. Records sent to server every few seconds (or more) from their mobile devices.
- Profile : 62% male, 38% female ; a concentration of young-age population
- Data : Demographic (very limited) ; Phone GPS data, Wi-Fi Scans, accelerometer (sum up to about 50 million unique location points).
- Management : Expansion up to 400 GB via Spark, a resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines.

Research Questions Boiled Down

Given those data, there are several research questions we want to address :

- Q1 : (Mobility) Is there a general suggestion of when does a person's mobility pattern "converges" ?

Research Questions Boiled Down

Given those data, there are several research questions we want to address :

- Q1 : (Mobility) Is there a general suggestion of when does a person's mobility pattern "converges" ?
- – This recommendation is likely the minimum time we seek.

Research Questions Boiled Down

Given those data, there are several research questions we want to address :

- Q1 : (Mobility) Is there a general suggestion of when does a person's mobility pattern "converges" ?
- – This recommendation is likely the minimum time we seek.
- Q2 : (Activity Space) How much difference in terms of coverage on activity space do we have for a data collection period of 1/7/30 days versus our recommendation ?

Research Questions Boiled Down

Given those data, there are several research questions we want to address :

- Q1 : (Mobility) Is there a general suggestion of when does a person's mobility pattern "converges" ?
- – This recommendation is likely the minimum time we seek.
- Q2 : (Activity Space) How much difference in terms of coverage on activity space do we have for a data collection period of 1/7/30 days versus our recommendation ?
- Q3 : (Exposure) How much difference in terms of predictive accuracy on the use of time budget do we have for a data collection period versus our recommendation ?

Research Questions Boiled Down

Given those data, there are several research questions we want to address :

- Q1 : (Mobility) Is there a general suggestion of when does a person's mobility pattern "converges" ?
- – This recommendation is likely the minimum time we seek.
- Q2 : (Activity Space) How much difference in terms of coverage on activity space do we have for a data collection period of 1/7/30 days versus our recommendation ?
- Q3 : (Exposure) How much difference in terms of predictive accuracy on the use of time budget do we have for a data collection period versus our recommendation ?
- – Q2 and Q3 also serve as a verification of our answers to Q1.

How do we define 'Convergence'?

For research on human mobility and activity space, convergence can be considered the state when a relatively stable and predictable spatio-temporal pattern is observed.

How do we define 'Convergence'?

For research on human mobility and activity space, convergence can be considered the state when a relatively stable and predictable spatio-temporal pattern is observed. In our study, we use the term 'last crossing time' (LCT) to measure it.

How do we define 'Convergence'?

For research on human mobility and activity space, convergence can be considered the state when a relatively stable and predictable spatio-temporal pattern is observed. In our study, we use the term 'last crossing time' (LCT) to measure it.

Last Crossing Time

Given the mean speed of person i 's travel behavior up to time t_i as $\mu(t_i)$, and the total observation time for this individual i as T_i, the last crossing time $\tilde{t}_i$ for this person i is the time s.t. the mean speed up to this time never exits the region within a tolerance bound of δ :

$$\arg\max \tilde{t}_i := \{\max t_i \mid \mu(\hat{t}_i) \in (\mu(T_i) \pm \delta\mu(T_i)) \ \forall \ t_i \leq \hat{t}_i \leq T_i\}$$

How do we define 'Convergence'?

For research on human mobility and activity space, convergence can be considered the state when a relatively stable and predictable spatio-temporal pattern is observed. In our study, we use the term 'last crossing time' (LCT) to measure it.

Last Crossing Time

Given the mean speed of person i 's travel behavior up to time t_i as $\mu(t_i)$, and the total observation time for this individual i as T_i, the last crossing time $\tilde{t}_i$ for this person i is the time s.t. the mean speed up to this time never exits the region within a tolerance bound of δ :

$$\arg\max \tilde{t}_i := \{\max t_i \mid \mu(\hat{t}_i) \in (\mu(T_i) \pm \delta\mu(T_i)) \ \forall \ t_i \leq \hat{t}_i \leq T_i\}$$

Notes : Mean Speed by time t : $\mu(t) = \frac{\sum \Delta d}{\sum \Delta t}$

Example

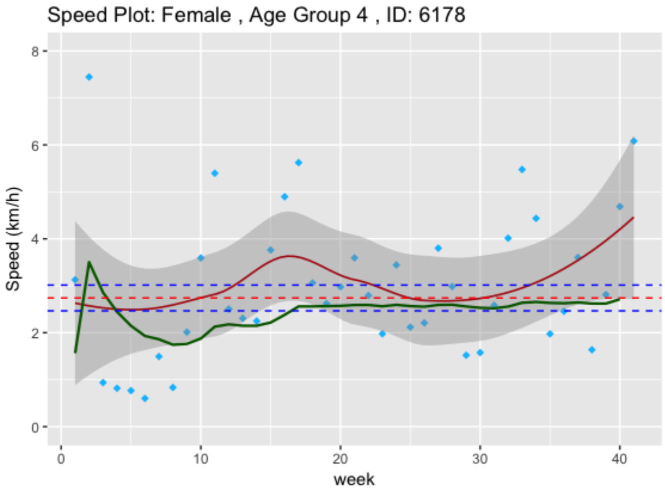


FIGURE – An Example of Convergence

Distribution

Here is the distribution of the 185 participants w.r.t. to their last crossing time using 10% tolerance.

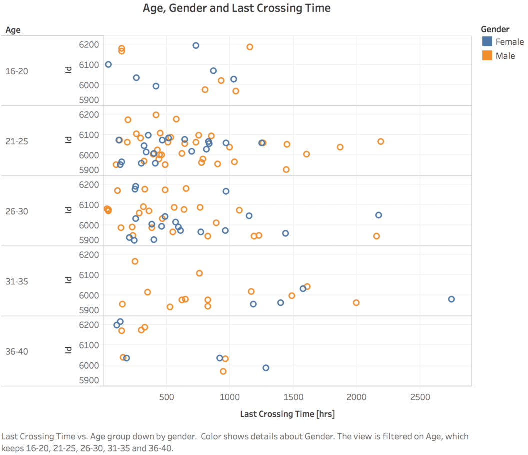


FIGURE – LCT w.r.t. Age/Gender Group

Is this robust ?

Can we see this outcome as reliable ?

Is this robust ?

Can we see this outcome as reliable ?
No !

Is this robust ?

Can we see this outcome as reliable ?
No ! There is a list of hazards :

- Outliers (Again, you fly to Sri Lanka)
- Nonrandomness in frequency of observation due to seasonality
- Too large time gap (for some, that might be more than few days)

Is this robust ?

Can we see this outcome as reliable ?
No ! There is a list of hazards :

- Outliers (Again, you fly to Sri Lanka)
- Nonrandomness in frequency of observation due to seasonality
- Too large time gap (for some, that might be more than few days)

To deal with this, we use a rebuilder (a.k.a. 'build-a-new-week') algorithm.

Nonrandomness in Observations I

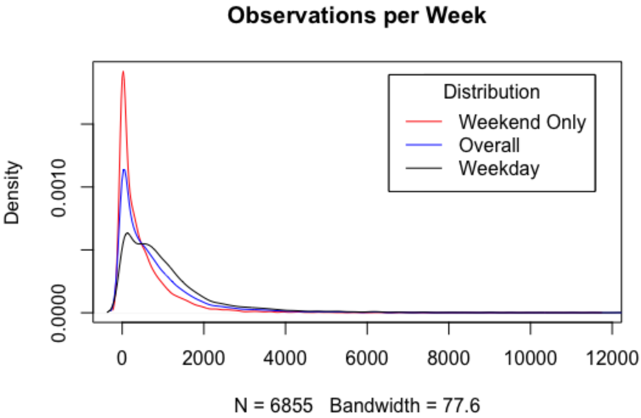


FIGURE – KDE on Weekday vs. Weekend

Nonrandomness in Observations II

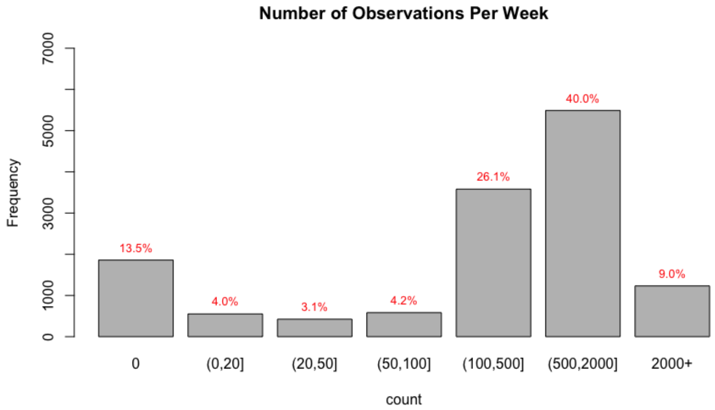


FIGURE – Bar Plot of Observations on Different Day-of-Week

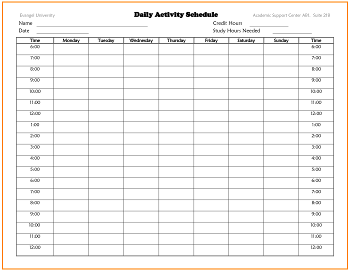
'Build a new week' algorithm

By essence, we consider simulating a person's new week by also considering the 'hour-of-day' and 'day-of-week' variety, because they are not necessarily distributed evenly.

'Build a new week' algorithm

By essence, we consider simulating a person's new week by also considering the 'hour-of-day' and 'day-of-week' variety, because they are not necessarily distributed evenly.

We sample with replacement from 'travel itineraries' created by a person's travel behaviors on Monday, until it filled up the three-hour block (e.g. 6AM-9AM). Then, we repeat this for Tuesday, Wednesday, ... Sunday.


Evangel University

Daily Activity Schedule

Academic Support Center AB1, Suite 218

Name ____________ Credit Hours ____________

Date ____________ Study Hours Needed ____________

Time	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Time
6:00								6:00
7:00								7:00
8:00								8:00
9:00								9:00
10:00								10:00
11:00								11:00
12:00								12:00
1:00								1:00
2:00								2:00
3:00								3:00
4:00								4:00
5:00								5:00
6:00								6:00
7:00								7:00
8:00								8:00
9:00								9:00
10:00								10:00
11:00								11:00
12:00								12:00

Does Gender Play a Big Role Here ?

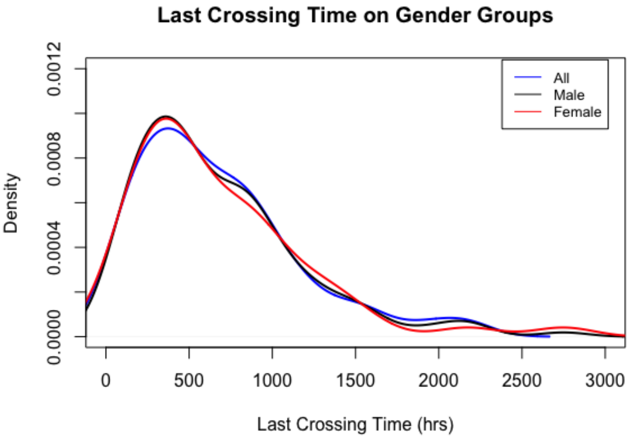


FIGURE – LCT By Gender After Algorithm

Does Age Play a Big Role Here ?

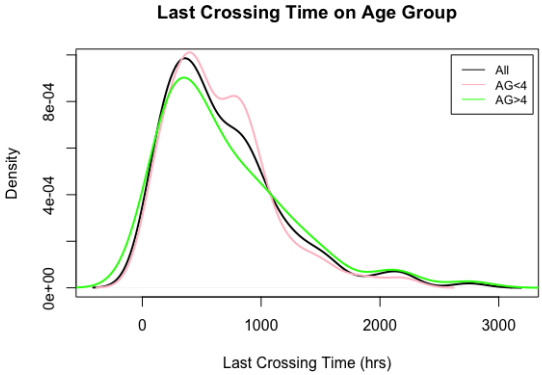


FIGURE – LCT By Age After Algorithm

Does Age Play a Big Role Here ?

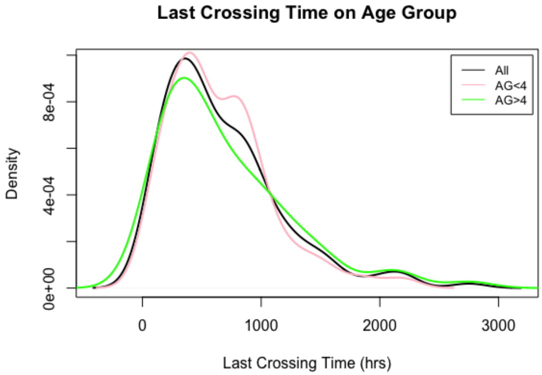


FIGURE – LCT By Age After Algorithm

More specified age group shows weak differences.

Take-Home Message

1. We found weak variation on stability of mobility across age/gender groups.

Take-Home Message

1. We found weak variation on stability of mobility across age/gender groups.
2. By Shapiro-Wilks Normality test, the distribution of LCT is normal across different age groups, lesser for gender.

Take-Home Message

1. We found weak variation on stability of mobility across age/gender groups.
2. By Shapiro-Wilks Normality test, the distribution of LCT is normal across different age groups, lesser for gender.

Here is the proportion of last crossing time observed by different lengths of sensor data collection with a 10% of tolerance bound δ :

Sensor Length (Days)	1	3	7	**14**	21	30
LCT Observed	8%	19%	**45%**	**83%**	90%	94%

Take-Home Message

1. We found weak variation on stability of mobility across age/gender groups.
2. By Shapiro-Wilks Normality test, the distribution of LCT is normal across different age groups, lesser for gender.
Here is the proportion of last crossing time observed by different lengths of sensor data collection with a 10% of tolerance bound δ :

Sensor Length (Days)	1	3	7	**14**	21	30
LCT Observed	8%	19%	**45%**	**83%**	90%	94%

3. Generally, researchers should not adopt a 'golden standard' of how many days of sensor data researchers should collect without context. We have a general recommendation of 14 days versus the 7-day convention for most health research.

From KDE to Density Ranking Algorithm (I)

(Part adapted from Chen (2017) CS&SS talk)

Given a collection of points, a common statistical approach is to estimate the probability density function (PDF). Based on the estimated density function, we can then compare these datasets. A popular and simple approach called the kernel density estimation (KDE), which is calculated as belows :

From KDE to Density Ranking Algorithm (I)
(Part adapted from Chen (2017) CS&SS talk)

Given a collection of points, a common statistical approach is to estimate the probability density function (PDF). Based on the estimated density function, we can then compare these datasets. A popular and simple approach called the kernel density estimation (KDE), which is calculated as belows :

$$\hat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{X_i - x}{h}),$$

From KDE to Density Ranking Algorithm (I)
(Part adapted from Chen (2017) CS&SS talk)

Given a collection of points, a common statistical approach is to estimate the probability density function (PDF). Based on the estimated density function, we can then compare these datasets. A popular and simple approach called the kernel density estimation (KDE), which is calculated as belows :

$$\hat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{X_i - x}{h}),$$

where $K(\cdot)$ is the kernel function that is often a smooth function like a Gaussian, and $h > 0$ is the smoothing bandwidth that controls the amount of smoothing.

From KDE to Density Ranking Algorithm (I)

(Part adapted from Chen (2017) CS&SS talk)

Given a collection of points, a common statistical approach is to estimate the probability density function (PDF). Based on the estimated density function, we can then compare these datasets. A popular and simple approach called the kernel density estimation (KDE), which is calculated as belows :

$$\hat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{X_i - x}{h}),$$

where $K(\cdot)$ is the kernel function that is often a smooth function like a Gaussian, and $h > 0$ is the smoothing bandwidth that controls the amount of smoothing.
Problems ?

From KDE to Density Ranking Algorithm (I)
(Part adapted from Chen (2017) CS&SS talk)

Given a collection of points, a common statistical approach is to estimate the probability density function (PDF). Based on the estimated density function, we can then compare these datasets. A popular and simple approach called the kernel density estimation (KDE), which is calculated as belows :

$$\hat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{X_i - x}{h}),$$

where $K(\cdot)$ is the kernel function that is often a smooth function like a Gaussian, and $h > 0$ is the smoothing bandwidth that controls the amount of smoothing.

Problems ?

The KDE cannot detect intricate structures inside the GPS data because the underlying PDF does not exist, hence our probability distribution function is singular.

From KDE to Density Ranking Algorithm (II)

(Part adapted from Chen (2017) CS&SS talk)

The density ranking (Chen 2016 ; Chen and Dobra 2017) is a transformed quantity/function from the KDE. The main idea is that, instead of using the density value, we focus on the ranking of it.

From KDE to Density Ranking Algorithm (II)
(Part adapted from Chen (2017) CS&SS talk)

The density ranking (Chen 2016 ; Chen and Dobra 2017) is a transformed quantity/function from the KDE. The main idea is that, instead of using the density value, we focus on the ranking of it.
The density ranking at point x is

$$\hat{\alpha}(x) = \frac{1}{n} \sum_{i=1}^{n} I(\hat{p}(x) \geq \hat{p}(X_i))$$

From KDE to Density Ranking Algorithm (II)
(Part adapted from Chen (2017) CS&SS talk)

The density ranking (Chen 2016 ; Chen and Dobra 2017) is a transformed quantity/function from the KDE. The main idea is that, instead of using the density value, we focus on the ranking of it.
The density ranking at point x is

$$\hat{\alpha}(x) = \frac{1}{n} \sum_{i=1}^{n} I(\hat{p}(x) \geq \hat{p}(X_i))$$

To compare multiple density rankings from multiple datasets, a simple approach is to overlap level plots. For a density ranking $\hat{\alpha}$, let

$$\hat{A}_\gamma = \{x : \hat{\alpha}(x) \geq 1 - \gamma\}$$

be the (upper) level set, hence compare the density ranking of each individual by overlapping their level sets at different levels.

From KDE to Density Ranking Algorithm (III)
(Part adapted from Chen (2017) CS&SS talk)

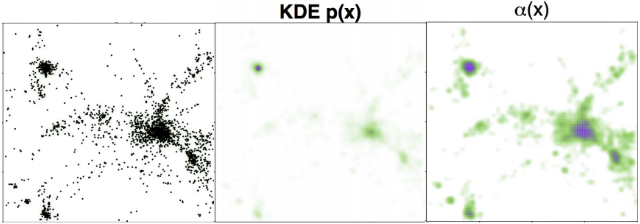


FIGURE – KDE vs Density Ranking

From KDE to Density Ranking Algorithm (IV)
(Part adapted from Chen (2017) CS&SS talk)

As a summary, density ranking methods :

- Density ranking $\hat{\alpha}(x)$ can be viewed as an estimator to certain characteristics of the underlying population distribution ; It also converges to $\alpha(x)$ in topological sense ;

From KDE to Density Ranking Algorithm (IV)
(Part adapted from Chen (2017) CS&SS talk)

As a summary, density ranking methods :

- Density ranking $\hat{\alpha}(x)$ can be viewed as an estimator to certain characteristics of the underlying population distribution ; It also converges to $\alpha(x)$ in topological sense ;
- Density ranking is still a consistent estimator even when the density *does not exist* !

From KDE to Density Ranking Algorithm (IV)

(Part adapted from Chen (2017) CS&SS talk)

As a summary, density ranking methods :

- Density ranking $\hat{\alpha}(x)$ can be viewed as an estimator to certain characteristics of the underlying population distribution ; It also converges to $\alpha(x)$ in topological sense ;
- Density ranking is still a consistent estimator even when the density *does not exist* !
- The population density ranking to a singular measure can be generalized by the concept of the *Hausdorff (geometric) density*

From KDE to Density Ranking Algorithm (IV)
(Part adapted from Chen (2017) CS&SS talk)

As a summary, density ranking methods :

- Density ranking $\hat{\alpha}(x)$ can be viewed as an estimator to certain characteristics of the underlying population distribution ; It also converges to $\alpha(x)$ in topological sense ;
- Density ranking is still a consistent estimator even when the density *does not exist* !
- The population density ranking to a singular measure can be generalized by the concept of the *Hausdorff (geometric) density*
- To verify our recommendation in Part 3, we examine the coverage by hierarchical clustering, and found our recommendation improved on coverage by over 60% compared to the ones with right truncation at Day 7.

From KDE to Density Ranking Algorithm (III)
(Part adapted from Chen (2017) CS&SS talk)

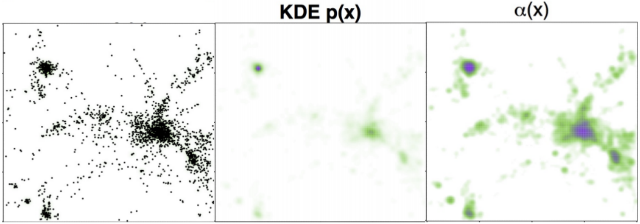


FIGURE – KDE vs Density Ranking

Coming Soon : seq2seq model

For predictive models on coverage, we can use a deep learning method called seq2seq model, for which we draw a multi-layer sequence-to-sequence network with LSTM cells and attention mechanism in the decoder looks like this.

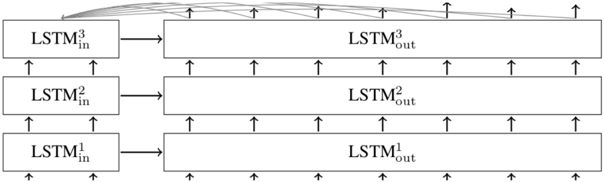


FIGURE – seq2seq models

Future Work

Our work is the first empirical attempt to explore the relationship between the length of sensor data collection per individual and its effect on the size of "information set' (sample) we have, as well as its implications on the application of health research. In the future, our work will consider :

Future Work

Our work is the first empirical attempt to explore the relationship between the length of sensor data collection per individual and its effect on the size of "information set' (sample) we have, as well as its implications on the application of health research. In the future, our work will consider :

- The choice of δ for 'convergence' may be better defined.
- – is $\delta = 0.1$ a good tolerance bound ?

Future Work

Our work is the first empirical attempt to explore the relationship between the length of sensor data collection per individual and its effect on the size of "information set' (sample) we have, as well as its implications on the application of health research. In the future, our work will consider :

- The choice of δ for 'convergence' may be better defined.
- – is $\delta = 0.1$ a good tolerance bound ?
- More demographic information could be added as covariates

Future Work

Our work is the first empirical attempt to explore the relationship between the length of sensor data collection per individual and its effect on the size of "information set' (sample) we have, as well as its implications on the application of health research. In the future, our work will consider :

- The choice of δ for 'convergence' may be better defined.
- – is $\delta = 0.1$ a good tolerance bound ?
- More demographic information could be added as covariates
- How can we use such data to measure social interaction ? (TDN)

Future Work

Our work is the first empirical attempt to explore the relationship between the length of sensor data collection per individual and its effect on the size of "information set' (sample) we have, as well as its implications on the application of health research. In the future, our work will consider :

- The choice of δ for 'convergence' may be better defined.
- – is $\delta = 0.1$ a good tolerance bound ?
- More demographic information could be added as covariates
- How can we use such data to measure social interaction ? (TDN)
- Our experiments are limited by the number of participants. How do we collect such data from a more diverse group of people ?

Thank you !

We thank Prof. Kyle Crowder from Dept. of Sociology at the University of Washington with initial comments on how this work could apply to ongoing research questions in demography.