Lessons from Big Data
— What can Pedometrics Learn

Jianghao Wang
State Key Laboratory of Resources and Environmental Information System,
Institute of Geographic Sciences and Natural Resources Research,
Chinese Academy of Sciences

June 28 2017 @ Pedometrics 2017
1. Introduction
2. Big Data in Urban Studies
3. Lessons from Big Data
Nowadays, our world sustainable development faces significant challenges.

Big data, from earth observation system to our daily lift, providing a new way to better understanding our world.

Data is the new oil of the digital world.
Traditional Way to Model Our World

- Geographical Information System (GIS)
- Global positioning System (GPS)
- Remote Sensing (RS)
- Sample survey
- Statistics

Geographical Information System (GIS)

Global positioning System (GPS)
New Technologies to Understand Our World

Cloud Computing
( Brain )

A

Big data / Human being

B

Artificial Intelligence
( thinking mode )

C

Mobile Internet
( neural network transmission )

D

IoT / Sensor Network
( Sensory organs )

E

Social Media
( social way in cyberspace )
2 Big Data in Urban Studies
Pan-space Urban Sensing is a crossover study investigating the use of digital sensors (including remote sensing, internet of things, wireless sensor network, citizen-sensing, et al.) to obtain, compute, and analyze urban dynamic environments to achieve more sustainable and efficient cities.
Pan-Space Urban Sensing

**Physical Urban Space**
Sensing urban physical structure with remote sensing techniques

1

**Human Activity Space**
Sensing human activities or behaviors with spatio-temporal big data

2

**Social-economic Space**
Sensing urban social-economic with census, IoT, and electronic business

3

**Cyber Urban Space**
Sensing urban in cyberspace with social media and mobile network

4
Pan-Space Urban Sensing

Pan-space Urban Sensing Technology

Physical Urban Space

Remote sensing

Human Activity Space

Lidar photogrammetry

Social-economic Space

Street View mapping system

Cyber Urban Space

Sensing urban physical structure with remote sensing techniques
To get the basic information about our city, like land use/land cover, building, roads, greens, etc.
Pan-Space Urban Sensing

Pan-space Urban Sensing Technology

Physical Urban Space

Human Activity Space

Social-economic Space

Cyber Urban Space

Sensing human activities or behaviors with spatio-temporal big data
Understanding human dynamic distribution, mobility, and living behaviors with mobile phone or smart card data.
Sensing urban social-economic with census, IoT, and electronic business data
Using social-economic related geospatial big data to help better understanding our social and economic system.
Pan-Space Urban Sensing

**Pan-space Urban Sensing Technology**

- **Physical Urban Space**
- **Human Activity Space**
- **Social-economic Space**
- **Cyber Urban Space**

**Internet social network**

**Human emotions in Twitter**

**Sensing urban in cyberspace with social media and mobile network**

Understanding citizen behavior or urban environment through online social media and study the relationship between physical and cyber urban space.
Research on Urban Sensing

Pan-Space Urban Sensing Research

Physical Urban Space

Human Activity Space

Social-economic Space

Cyber Urban Space

Model Dynamic Urban Structure

Sensing Chinese Dynamics

Social Sensing Urban Vibrancy

Sensing Citizen Emotion and PM$_{2.5}$
Physical Urban Space

Model Dynamic Urban Structure
Urban Big Data Center

China Urban Big Data Center

**A pan-space, multi-scale, dynamic** China urban big database.

Data volume > 50 TB.

---

**Data Integration / Fusion**

Data covers main cities in China. Pan-space (physical, human, social-economic, and cyber space) data were integrated or fused, including but not limited to,

- Urban fundamental GIS data
- Census, Economic Census
- 365 * 24 hours MP positioning data
- Social media checkins ......
**Problem:** Remotely sensed imagery hard to tell the land use functions in urban area.

- Inferring urban detailed land use functions through pan-space urban information with **Tensor-based classification**
Tensor-based Land Use Classification

Tensor

Main advantages:
✓ Directly present the multiway info.
✓ Preserve the multi-dimensional correlations.

Objective function:

\[
L(U, V, W) = \left\| Y - \sum_{k=1}^{K} \Phi^{(k)} \circ \Psi^{(k)} \circ \Theta^{(k)} \right\|_F^2 + \Phi(U, V, W) \]

Respect to: \( U \geq 0, V \geq 0, W \geq 0 \)

Mixed high order spatial-temporal data

Working component + Living component + Entertainment component + Other component

Linear combination of various components
Inferring urban detailed land use functions through pan-space urban information with Tensor-based artificial intelligence algorithm.

- 20 classes
- Overall accuracy: 91%
Urban Cells

Understanding detailed urban structure and behave through the point-of-interesting (POI) data.

Each point with different colors represent a urban unit (shop, school, hospital, or factory, etc.)
Urban Pulse

Sensing and forecasting urban traffic congestion with GPS equipped floating vehicles data in urban arteries.

Algorithm: Bayesian deep learning with spatio-temporal correlations

Forecasting accuracy: 85%
Human Activity Space

Sensing Chinese Dynamics
**Dynamic Population Distribution**

Multiple ST data fusion

- Mobile Phone positioning
- Social media checkins

Geospatial big data + Census / Geographical data

- Land use
- RS night light index
- Census data

Hybrid high resolution dynamic population distribution model

\[ P_{st} = f[Z_m(x, y, n), \mu_1 \cdot LU, \mu_2 \cdot S, \mu_3 \cdot R, \mu_4 \cdot TN, ...] \]

Dynamic people distribution at multiple scale
Urban Rhythm

Using mobile phone positioning data, we map urban human dynamic at very fine spatio-temporal scales.
Dynamic population distribution in 100 m resolution was estimated by using census and multiple geospatial big data with hybrid area-to-point geostatistical technique.

Day time population (12:00)

Night time population (24:00)
Chinese Dynamics

Location Based Service (LBS) data from Tencent Big Data center.

800 million users / day
50 billion location request / day
Chinese Footprint

Mapping Chinese footprint worldwide with 400 millions Weibo (equate to Twitter) checkins to answer the question of “Where are the Chinese?”

Wang et al., 2015 EPA
The Chinese American estimation based on Weibo checkins was validated with US Census in county level.

**Corr = 0.905, P < 0.001**

Chinese American
(Based on 2010 US Census)

Chinese American
(Estimated based on Weibo checkins)
Understanding individual human mobility is of fundamental importance for many applications from urban planning to disease spreading and traffic forecasting.

We propose social media (Weibo) checkins as a proxy for human mobility, as it relies on publicly available data and provides high resolution positioning when users opt to geotag their posts with their current location.

### Data source
- Mobile phone data
- Social media Checkins
- Floating car
- Internet-of-thing

### Geocomputation
- Trajectory
- OD matrix

### Pattern extra.
- ST pattern
- Statistical mechanics

### Mobility model
- Mobility model
- Spatial Interaction

### Application
- Urban planning
- Traffic forecasting
- Disease spreading
- Human geography

\[ T_{ij} \simeq T_{ij}^{\text{grav}} = \frac{P_i^2 P_j^2}{d_{ij}^6} \]
Understanding Chinese Mobility from Weibo

- Chinese mobility patterns is measured by big data computational strategy for identifying hundreds of millions of individuals’ space–time footprint trajectories;
- We discovered dialect-based culture ties control the Chinese mobility pattern;
- Our study provides solid evidence that Weibo checkins can indeed be a useful proxy for tracking and predicting human movement.

\[
\log(T_{od} / L_o) = \beta_1 \cdot \log(\text{Commuting}_{od}) + \beta_2 \cdot \log(\text{Dialect}_{od}) + F_o + F_d + \text{controls} + \epsilon_{od}
\]
Understanding Chinese Mobility from Weibo

City-level Chinese mobility
origin-destination (OD) matrix

City-level Chinese mobility
in geographical space

Beijing
Shanghai
Chengdu
Shenzhen
Cross-validation

Floating population in 6th Census
Weibo estimated human mobility

Corr. = 0.85 (p-value < 0.001).
PM$_{2.5}$ Air Pollution in China

China’s high level of ambient air pollution causes sickness, excess mortality risk.

Study health impact --- Measure social cost?
Citizen Emotion Sensing and PM$_{2.5}$

Billions of geotagged Weibos 

Control Variables (Weather, event, income, city property etc.)

sentiment metric

spatial econometrics model

PM$_{2.5}$ ST concentration

$SENTIMENT_{it} = \alpha_0 + \alpha_1 PM2.5_{it} + \alpha_2 X_{it} + T_t + \gamma_t + \epsilon_{it}$

Main findings:

- One standard deviation increase in the PM2.5 concentration is associated with a 0.05-0.06 standard deviation decrease in the sentiment index.
- One standard deviation increase in the city’s PM2.5 concentration can be offset by a 6.5 thousand RMB ($940) increase in the city-level annual wage.

What a lovely weather! Let’s go camping. ⚽

Today is heavily air polluted. My nose is stopped up. 😞

Natural language machine learning
Emotion Sensing and Air Pollution

The Geography of Weibo posts, PM$_{2.5}$ concentration and sentiment index, and their national relationship.
What can Pedometrics Learn?

Lessons from Big Data
<table>
<thead>
<tr>
<th></th>
<th>Remote Sensing</th>
<th>Soil Sensing</th>
<th>Urban Sensing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organization</strong></td>
<td>Top-down ↓</td>
<td>↓ &amp; ↑</td>
<td>Bottom-up ↑</td>
</tr>
<tr>
<td><strong>Data accessibility</strong></td>
<td>☺</td>
<td>☹</td>
<td>☹</td>
</tr>
<tr>
<td><strong>Data representation</strong></td>
<td>☹</td>
<td>☺</td>
<td>☹</td>
</tr>
<tr>
<td><strong>Spatio-temporal</strong></td>
<td>☹</td>
<td>☺</td>
<td>☺</td>
</tr>
<tr>
<td><strong>Model independence</strong></td>
<td>☹</td>
<td>☹</td>
<td>☺</td>
</tr>
</tbody>
</table>
Small data and Big data

**Wide data** table with detailed attribute but limited samples

Optimized samples are unbiased but with large error-variance

Samples may biased but with small error-variance

Long data table with limited attribute but massive samples

Data fusion

BLUE estimator
Pedometrics Model with Big Data

Spatial statics ---\rightarrow space-time dynamic

Statistical model ---\rightarrow Hybrid or Data driven

Homogeneity ---\rightarrow Heterogeneity model

Normal

Power law

Long tail

(global \rightarrow local model)
• Separate model
  – Spatial analysis
  – Time series analysis

• Spatio-temporal model
  – Statistical model
  – Physical model for specific field
  – Statistical + physical model

• Spatio-temporal data driven model
  – Machine learning
  – Statistical deep learning
  – Artificial intelligence algorithms
Spatio-temporal Vary Coefficient Model

Considering the space-time heterogeneity and dependence simultaneously with hybrid varying coefficient model

\[ y_{i,t} = \sum_{k_1=1}^{p_1} x_{i,t,k_1} \beta_{k_1}(a_i, b_i, \tau_t) + \sum_{k_2=1}^{p_2} z_{i,t,k_2} \gamma_{k_2}(a_i, b_i) + \sum_{k_3=1}^{p_3} g_{i,t,k_3} \theta_{k_3}(\tau_t) + \sum_{k_4=1}^{p_4} h_{i,t,k_4} \zeta_{k_4} + \varepsilon_{i,t} \]

- **Spatio-temporal variation**
- **Pure spatial variation**
- **Pure temporal variation**
- **Constant effect**

\[ \gamma = \theta = \zeta = 0 \quad 1. \text{ Geographically and Temporally Weighted Regression (GTWR)} \]

\[ \beta = \theta = \zeta = 0 \quad 2. \text{ Geographically Weighted Regression (GWR)} \]

\[ \beta = \gamma = \zeta = 0 \quad 3. \text{ Time varying coefficient model (TWR)} \]

\[ \beta = \theta = 0 \quad 4. \text{ Semi spatially varying coefficient model (SVC)} \]

\[ \beta = \gamma = \theta = 0 \quad 5. \text{ linear regression model (OLS)} \]

Wu and Wang, 2017
Spatio-temporal Vary Coefficient Model

Explain the spatial-temporal varying relationship between two target variables.
“The big $n$ problem” in DSM: Modelling massive spatial or spatiotemporal datasets with (geo)statistical method often face with matrix inversion which requires about $O(n^3/3)$ flops.

**Solutions**

1. **Developing alternative computationally efficient strategy**, e.g.:
   - Integrated nested Laplace approximations (INLA, Rue et al. 2009)
   - Reduced Rank Models (Cressie et al. 2008);
   - The Nearest-Neighbor Gaussian Process (Datta et al. 2016)

2. **High performance computing**
   - Parallel computing, MPI, CODA, cloud computing
   - CPU – GPU – TPU (Google Tensor Processing Unit)
Big Data in Action

- **Acquisition**
  - Python
  - PHP
  - Java

- **Database**
  - PostGIS
  - CouchDB
  - Redis
  - Cassandra
  - MongoDB
  - Membase
  - Riak

- **Data Analysis**
  - OSGeo
  - R

- **Visualization**
  - MapBox
  - JavaScript

- **Domain Expertise**
- **Math & Statistics**
- **Computing skill**
Deep Thinking in Action

Big data + Big model → Big unsolved question

Cross-disciplinary cooperation!
Thanks

Q & A

Collaborators: Yong Ge, Siqi Zheng, Wenjie Wu, Xingjian Liu, et al.
Acknowledgement: The National Natural Science Foundation of China

http://jianghao.wang
wangjh@lreis.ac.cn