Spatial Data Science and BIG DATA Analytics for Geographic Information Sciences

Sudipto Banerjee

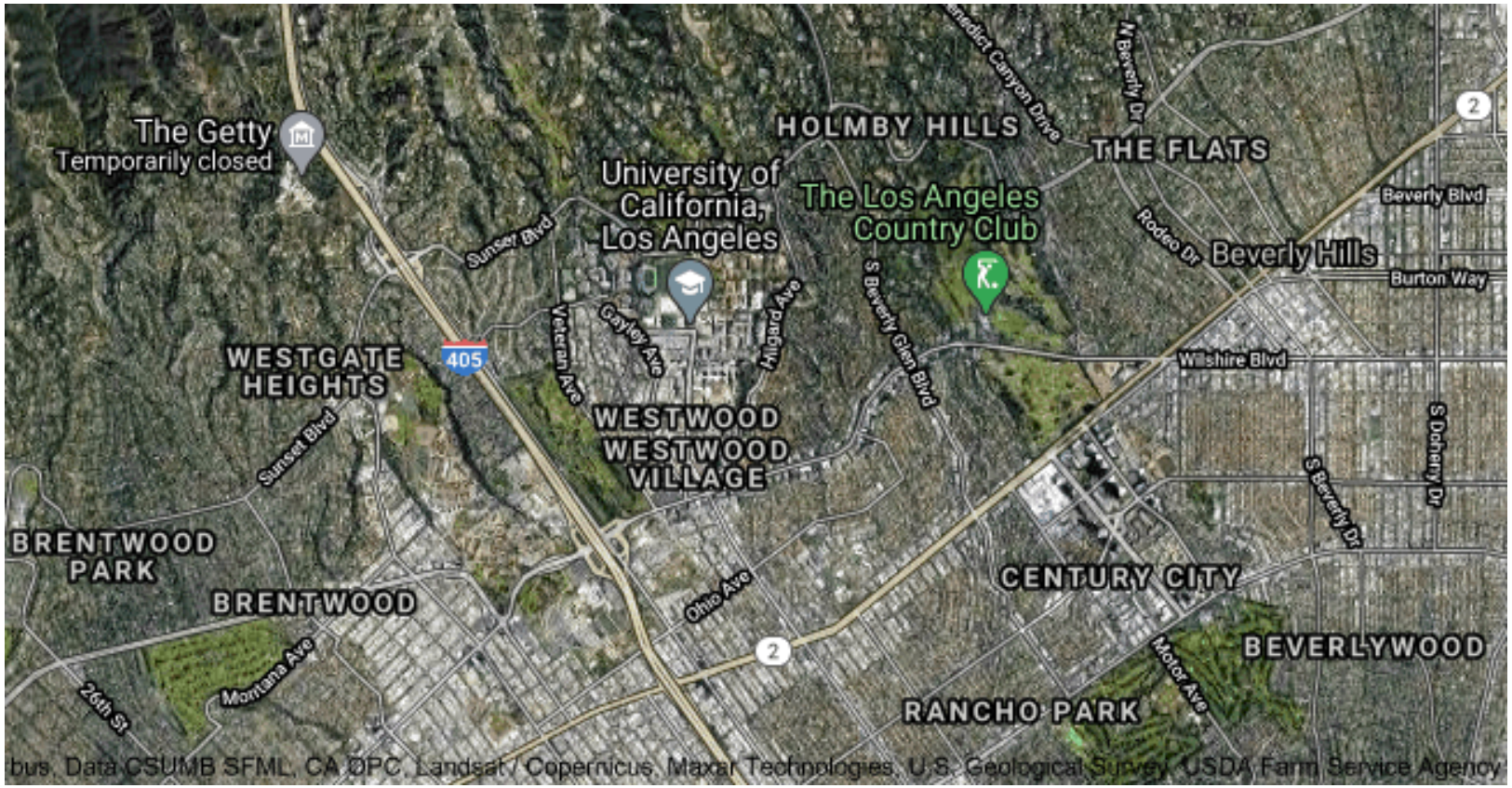
Professor and Chair of Biostatistics
Professor of Statistics
Affiliate Member, Institute of the Environment & Sustainability
University of California Los Angeles (UCLA)
sudipto@ucla.edu

Spring, 2021

UCLA FIELDING SCHOOL OF PUBLIC HEALTH

SIT: Spatial Information Technologies

Some neighbourhoods in Kolkata



Creating Databases for Maps

Global Positioning Systems (GPS)

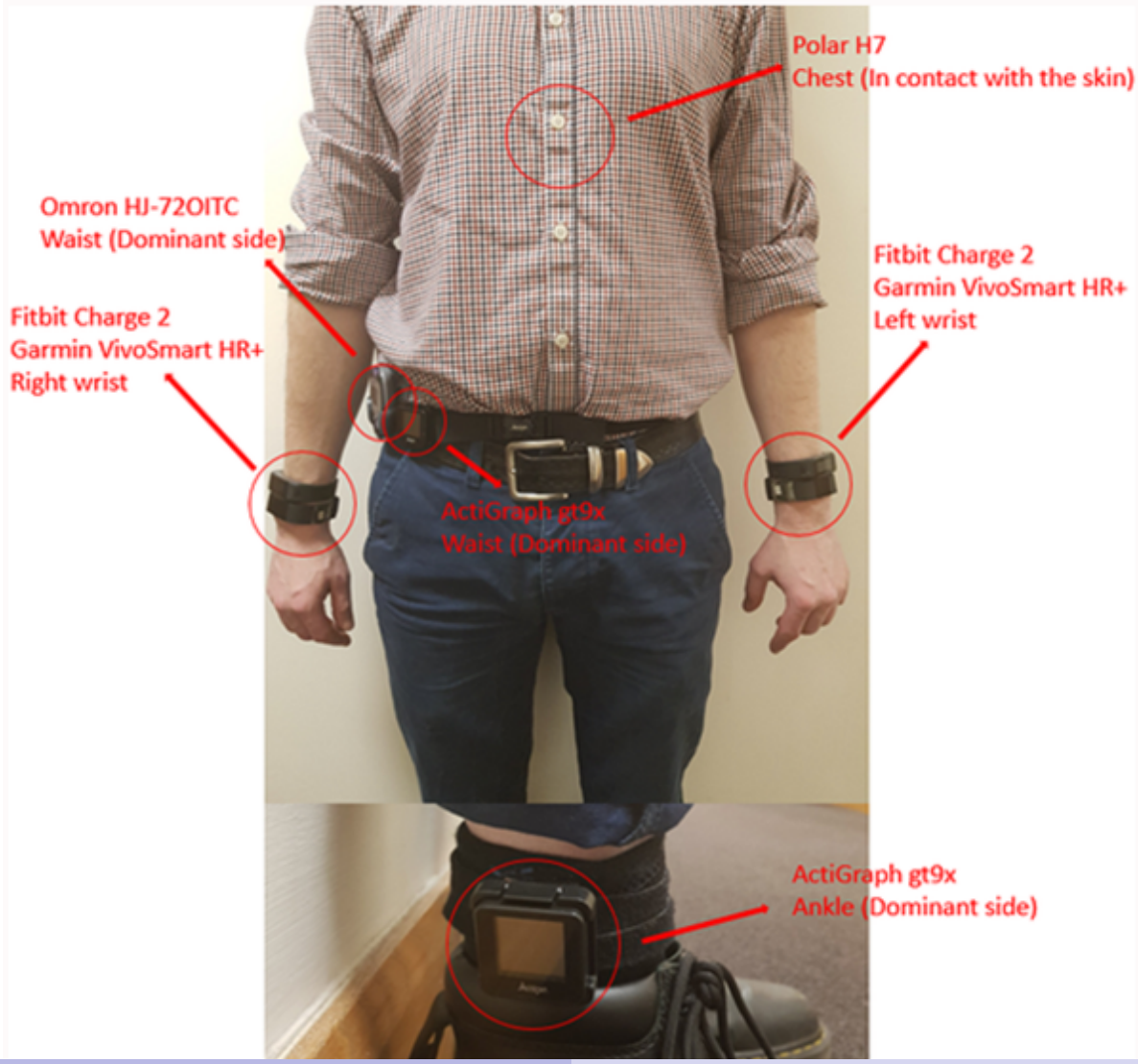
A wearable GPS device

Monitors for Physical Activity

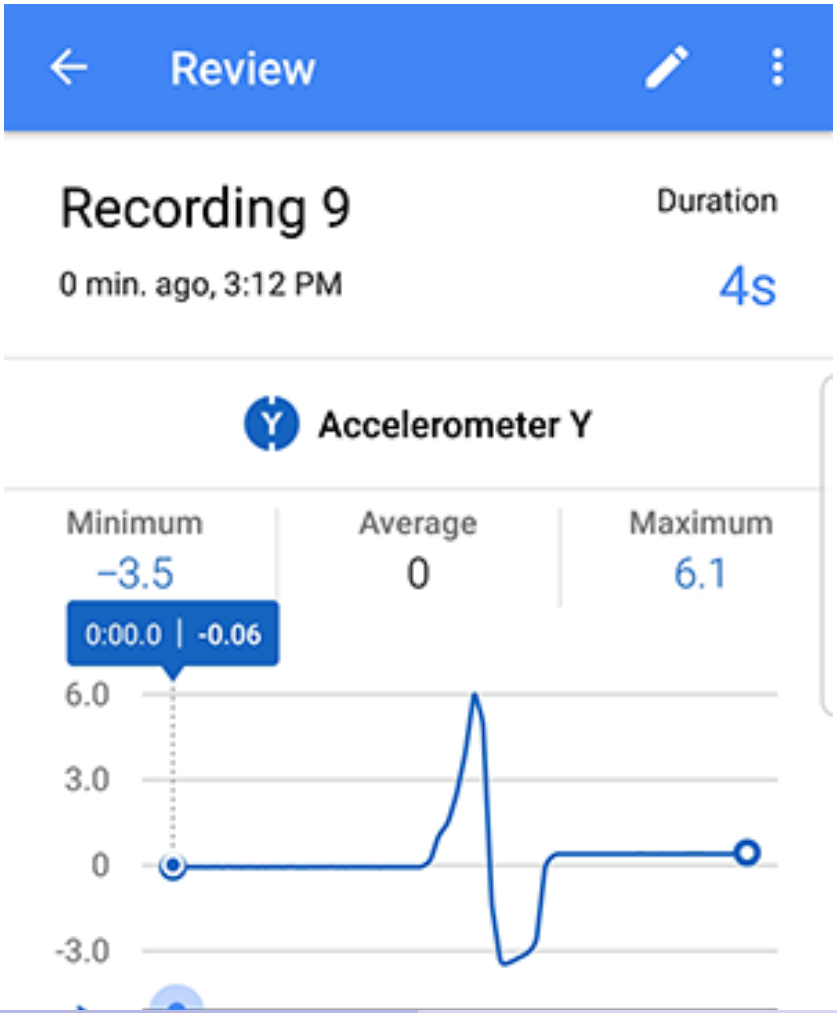
- Small motion sensor detectors (accelerometers) have generated substantial interest in monitoring human activity.
- Wearable devices, such as wrist-worn sensors that monitor gross motor activity (actigraph units) continuously record the activity levels of a subject, producing massive amounts of high-resolution measurements.

Actigraph units

Actigraph units (source: https://journals.plos.org/plosone/article/figures?id=10.1371/journal.pone.0216891)



Accelerometers: (source: https://www.sciencebuddies.org/science-fair-projects/references/accelerometer)



Accelerometers: From MAGs to METs

- Measure acceleration along multiple axes (2 or 3).

$$r(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}; \quad v(t) = \frac{d}{dt} r(t); \quad a(t) = \frac{d}{dt} v(t) = \begin{pmatrix} \ddot{x}(t) \\ \ddot{y}(t) \\ \ddot{z}(t) \end{pmatrix} .$$

- Accelerometers calculate $a(t)$ over small time-intervals
- Magnitude of acceleration:

$$MAG_{accel} = \sqrt{a_x^2 + a_y^2 + a_z^2},$$

where a_x, a_y and a_z are averaged accelerometer readings along the three axes over small (e.g., 10-sec) time windows (epochs).

- Metabolic Equivalent of Task (Sasaki, 2011):

$$\text{MET} = 0.000863 \cdot MAG + 0.668876$$

- With *MAG* readings from hip and ankle (Mortazavi et al., 2013):

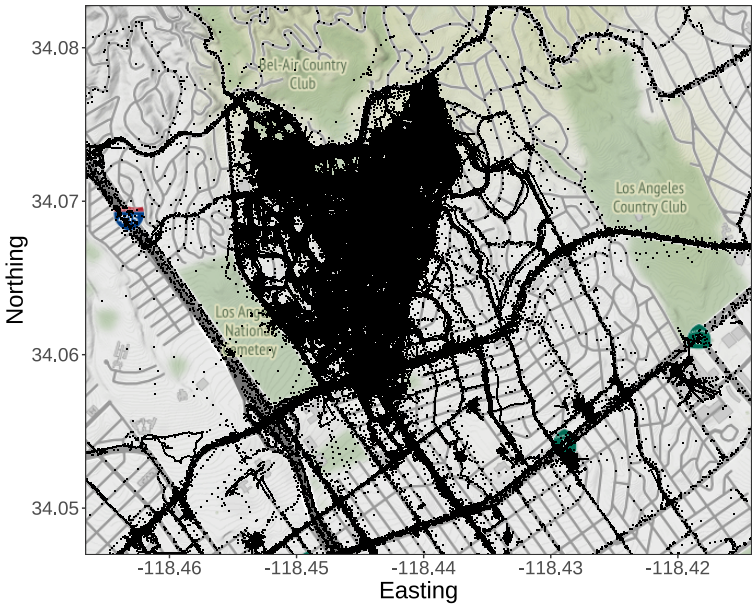
$$\text{MET} = 5.289 \cdot (MAG_{hip} + MAG_{ankle}) - 8.5548$$

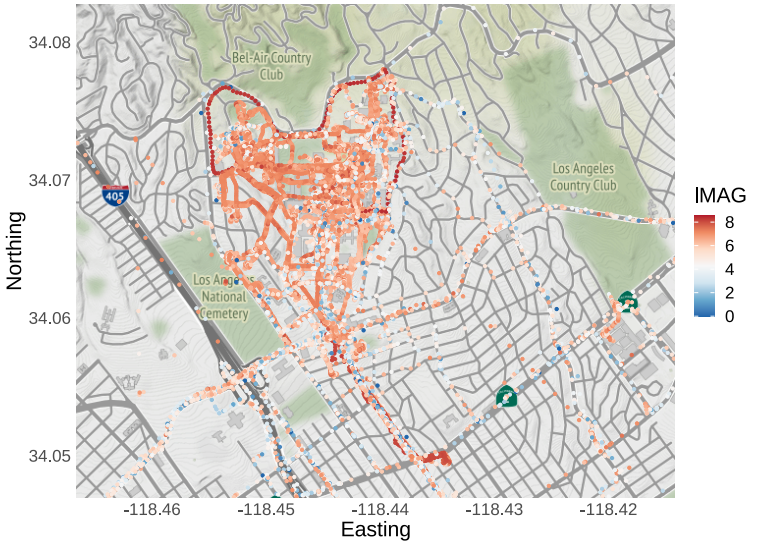
METs and MAGs by Activity Levels

Activity intensity	MET range	MAG
Sedentary or light	[0, 3)	[0, 493)
Moderate	[3, 6)	[493, 1029)
Hard	[6, 9)	[1029, 1608)
Very hard	[9, ∞)	[1608, ∞)

Table: MAG activity count cut-points for different PA intensity levels

Trajectories (from GPS) carved out by subjects



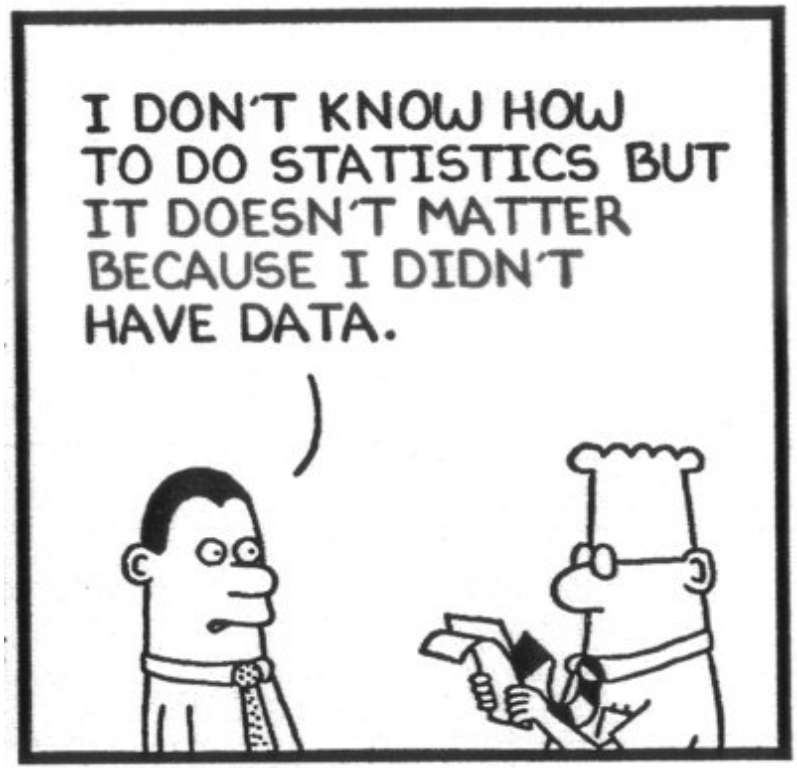


Statistics and Data Science...Is Happening

Hal Varian, Google's chief economist:

The ability to take data-to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it-that's going to be a hugely important skill in the next decades

Source: `http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286`



The Information Revolution

- We live in an Information Age.
- Computers collect and store information in quantities that were earlier unimaginable.
- What is this information?
  - Measurements, counts, costs, sales revenue...
  - arising in sciences, public health, business...
- Raw, "undigested" data stored on computer disks is useless unless we make sense of it.
- Statistics: the art and science of extracting meaning from seemingly incomprehensible data.
- Make good use of information to make sound decisions.

The Information Revolution

- We live in an Information Age.
- Computers collect and store information in quantities that were earlier unimaginable.
- What is this information?
  - Measurements, counts, costs, sales revenue...
  - arising in sciences, public health, business...
- Raw, "undigested" data stored on computer disks is useless unless we make sense of it.
- Statistics: the art and science of extracting meaning from seemingly incomprehensible data.
- Make good use of information to make sound decisions.

The Information Revolution

- We live in an Information Age.
- Computers collect and store information in quantities that were earlier unimaginable.
- What is this information?
  - Measurements, counts, costs, sales revenue...
  - arising in sciences, public health, business...
- Raw, "undigested" data stored on computer disks is useless unless we make sense of it.
- Statistics: the art and science of extracting meaning from seemingly incomprehensible data.
- Make good use of information to make sound decisions.

The Information Revolution

- We live in an Information Age.
- Computers collect and store information in quantities that were earlier unimaginable.
- What is this information?
  - Measurements, counts, costs, sales revenue...
  - arising in sciences, public health, business...
- Raw, "undigested" data stored on computer disks is useless unless we make sense of it.
- Statistics: the art and science of extracting meaning from seemingly incomprehensible data.
- Make good use of information to make sound decisions.

The Information Revolution

- We live in an Information Age.
- Computers collect and store information in quantities that were earlier unimaginable.
- What is this information?
  - Measurements, counts, costs, sales revenue...
  - arising in sciences, public health, business...
- Raw, "undigested" data stored on computer disks is useless unless we make sense of it.
- Statistics: the art and science of extracting meaning from seemingly incomprehensible data.
- Make good use of information to make sound decisions.

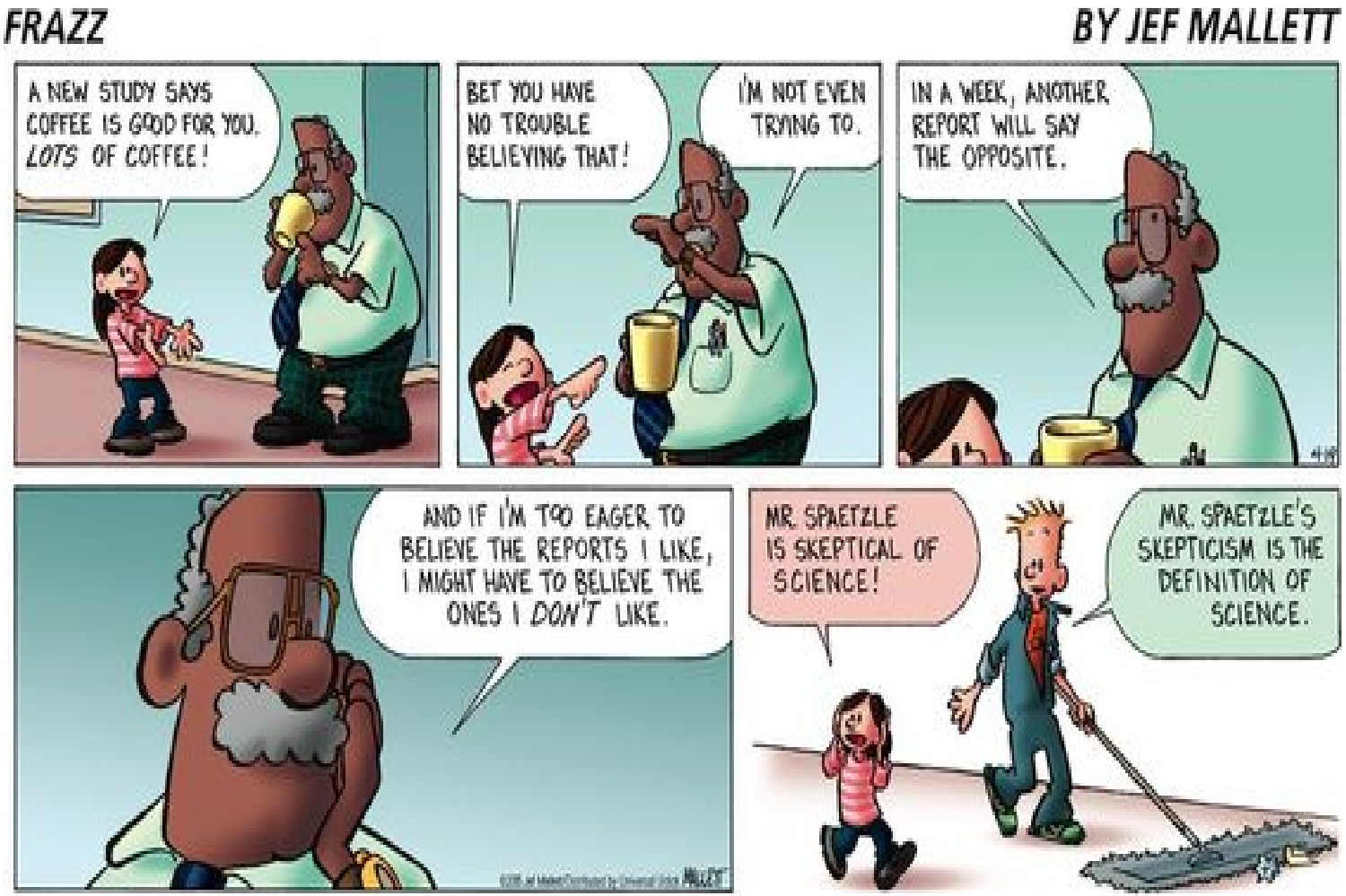
The Information Revolution

- We live in an Information Age.
- Computers collect and store information in quantities that were earlier unimaginable.
- What is this information?
  - Measurements, counts, costs, sales revenue...
  - arising in sciences, public health, business...
- Raw, “undigested” data stored on computer disks is useless unless we make sense of it.
- Statistics: the art and science of extracting meaning from seemingly incomprehensible data.
- Make good use of information to make sound decisions.

The Information Revolution

- We live in an Information Age.
- Computers collect and store information in quantities that were earlier unimaginable.
- What is this information?
  - Measurements, counts, costs, sales revenue...
  - arising in sciences, public health, business...
- Raw, “undigested” data stored on computer disks is useless unless we make sense of it.
- Statistics: the art and science of extracting meaning from seemingly incomprehensible data.
- Make good use of information to make sound decisions.

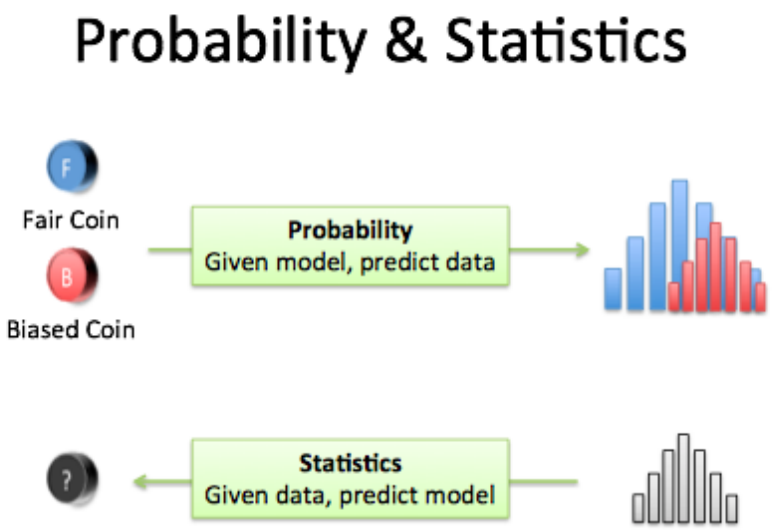
A favorite cartoon



What is Statistics?

Statistics uses math. Statistics is not math!

- Statistics: induction, inference, generalization.
- Statistics: From data, what can we conclude about the world?
- Mathematics: deduction.
- Math modeling: If this model is true, what data occurs?
- Problem: Data can occur under lots of models. Which model is right?
- Mathematics vital: Statistics uses math!



What is Statistics?

Statistics applies to almost any field of science

In one week a statistician may:

- help design an experiment to evaluate the effects of a new treatment for a disease,
- analyze data gathered in the Amazon rain forests by an ecologist,
- help predict how the California wild fires will spread,
- help Google design a better search engine,
- analyze data from researchers trying to find genes that cause breast cancer.

What is Statistics?

Statistics applies to almost any field of science

In one week a statistician may:

- help design an experiment to evaluate the effects of a new treatment for a disease,
- analyze data gathered in the Amazon rain forests by an ecologist,
- help predict how the California wild fires will spread,
- help Google design a better search engine,
- analyze data from researchers trying to find genes that cause breast cancer.

What is Statistics?

Statistics applies to almost any field of science

In one week a statistician may:

- help design an experiment to evaluate the effects of a new treatment for a disease,
- analyze data gathered in the Amazon rain forests by an ecologist,
- help predict how the California wild fires will spread,
- help Google design a better search engine,
- analyze data from researchers trying to find genes that cause breast cancer.

What is Statistics?

Statistics applies to almost any field of science

In one week a statistician may:

- help design an experiment to evaluate the effects of a new treatment for a disease,
- analyze data gathered in the Amazon rain forests by an ecologist,
- help predict how the California wild fires will spread,
- help Google design a better search engine,
- analyze data from researchers trying to find genes that cause breast cancer.

What is Statistics?

Statistics applies to almost any field of science

In one week a statistician may:

- help design an experiment to evaluate the effects of a new treatment for a disease,
- analyze data gathered in the Amazon rain forests by an ecologist,
- help predict how the California wild fires will spread,
- help Google design a better search engine,
- analyze data from researchers trying to find genes that cause breast cancer.

What is Statistics?

Statistics applies to almost any field of science

In one week a statistician may:

- help design an experiment to evaluate the effects of a new treatment for a disease,
- analyze data gathered in the Amazon rain forests by an ecologist,
- help predict how the California wild fires will spread,
- help Google design a better search engine,
- analyze data from researchers trying to find genes that cause breast cancer.

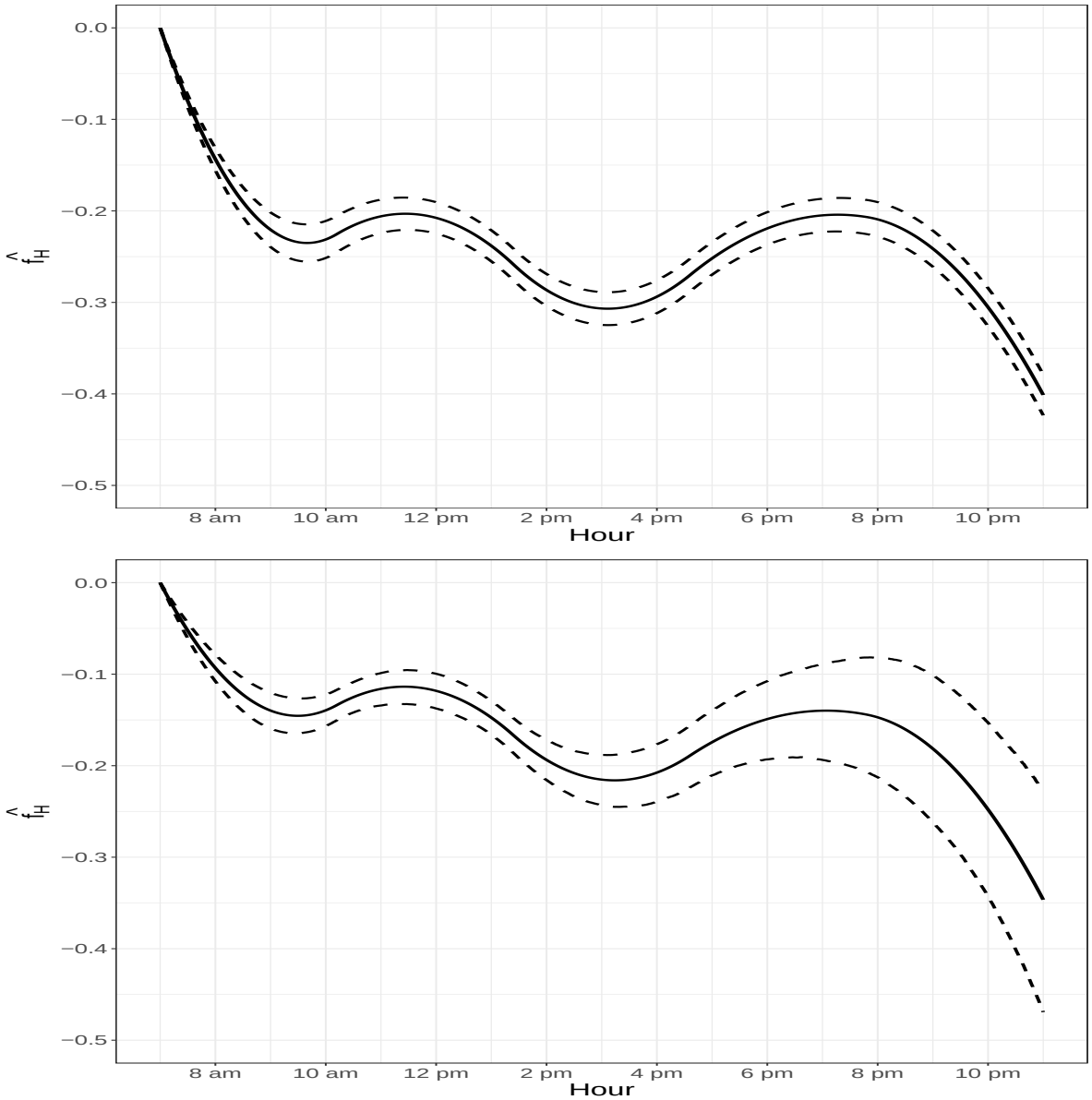
Bayesian Statistical Modeling and Inference

$$[\text{Unknown} \,|\, \text{Known}] = [\text{process}, \text{parameters} \,|\, \text{data}]$$

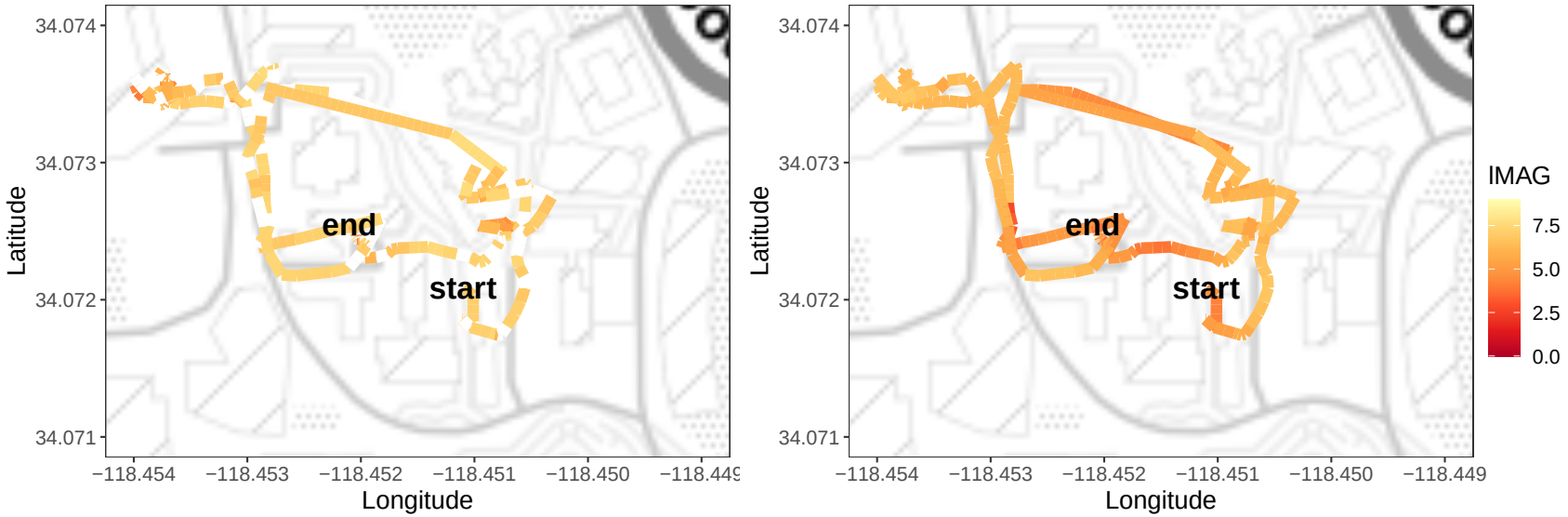
$$[\text{data} \,|\, \text{process}, \text{parameters}] \times [\text{process} \,|\, \text{parameters}] \times [\text{parameters}] \ .$$

- Process drives inference;
- Interpolate process at arbitrary locations;
- Predict scientific phenomenon (weather);
- Keep an eye on scalability (BIG DATA).

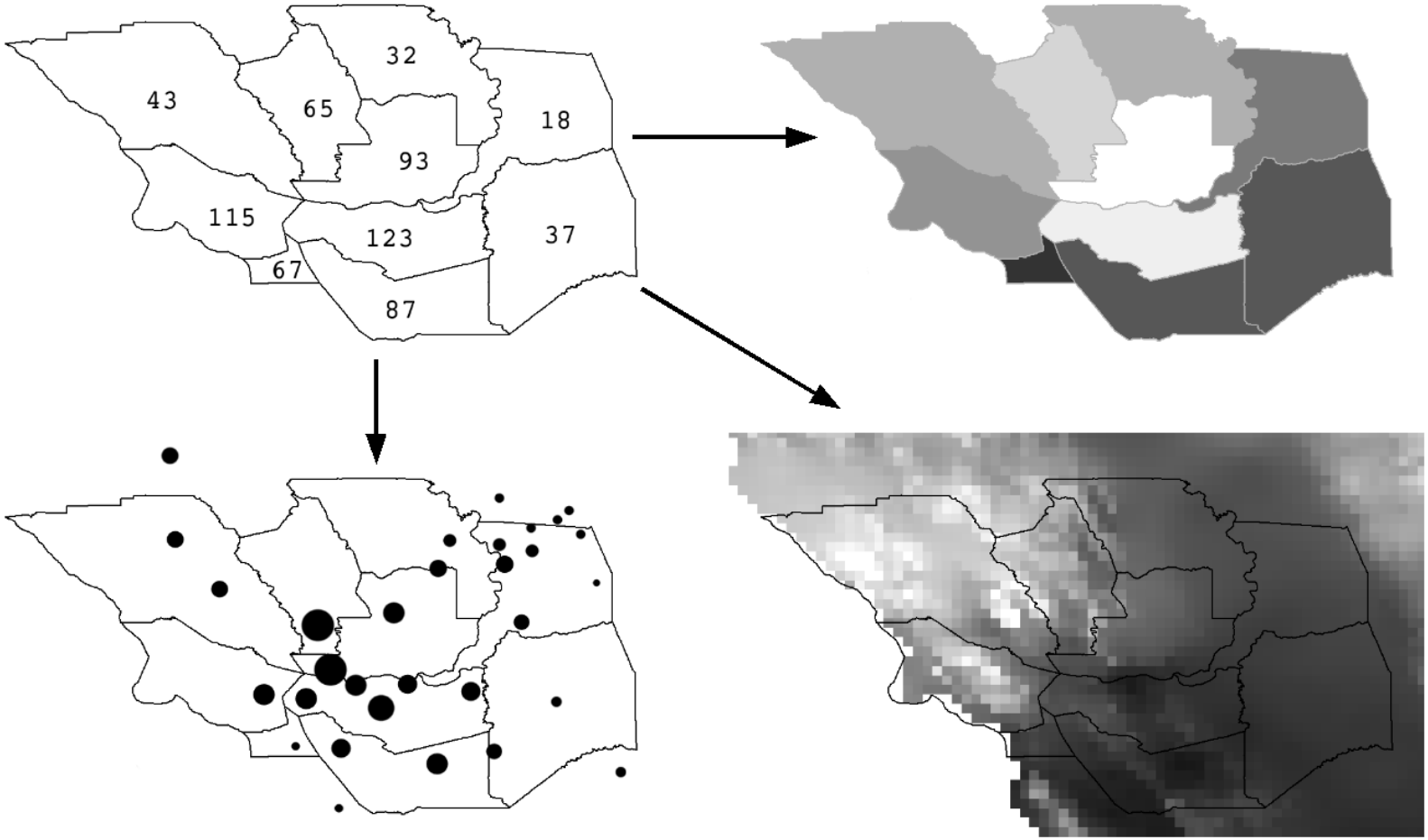
Estimating daily “periodic” behavior



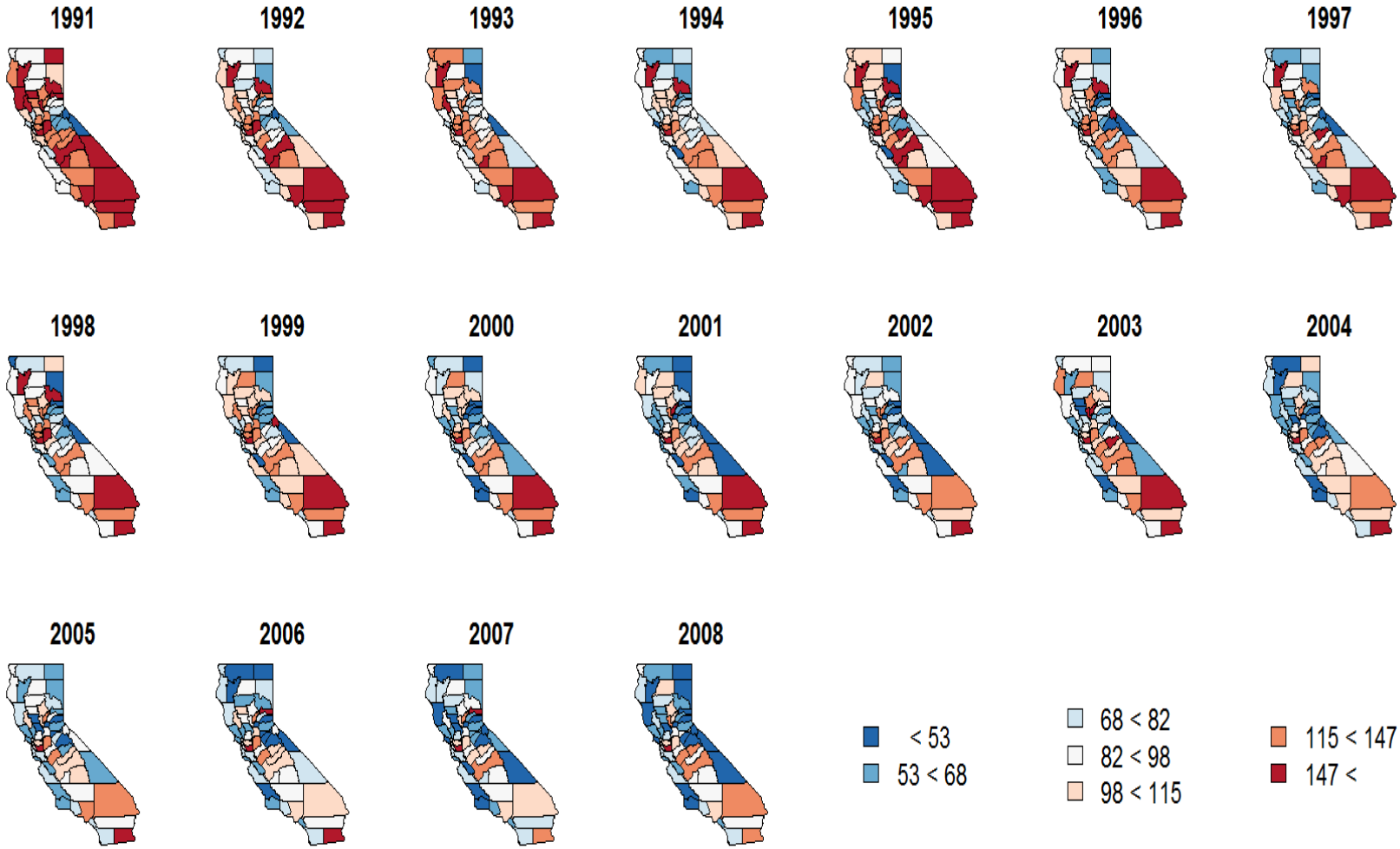
Predicted METs along trajectories



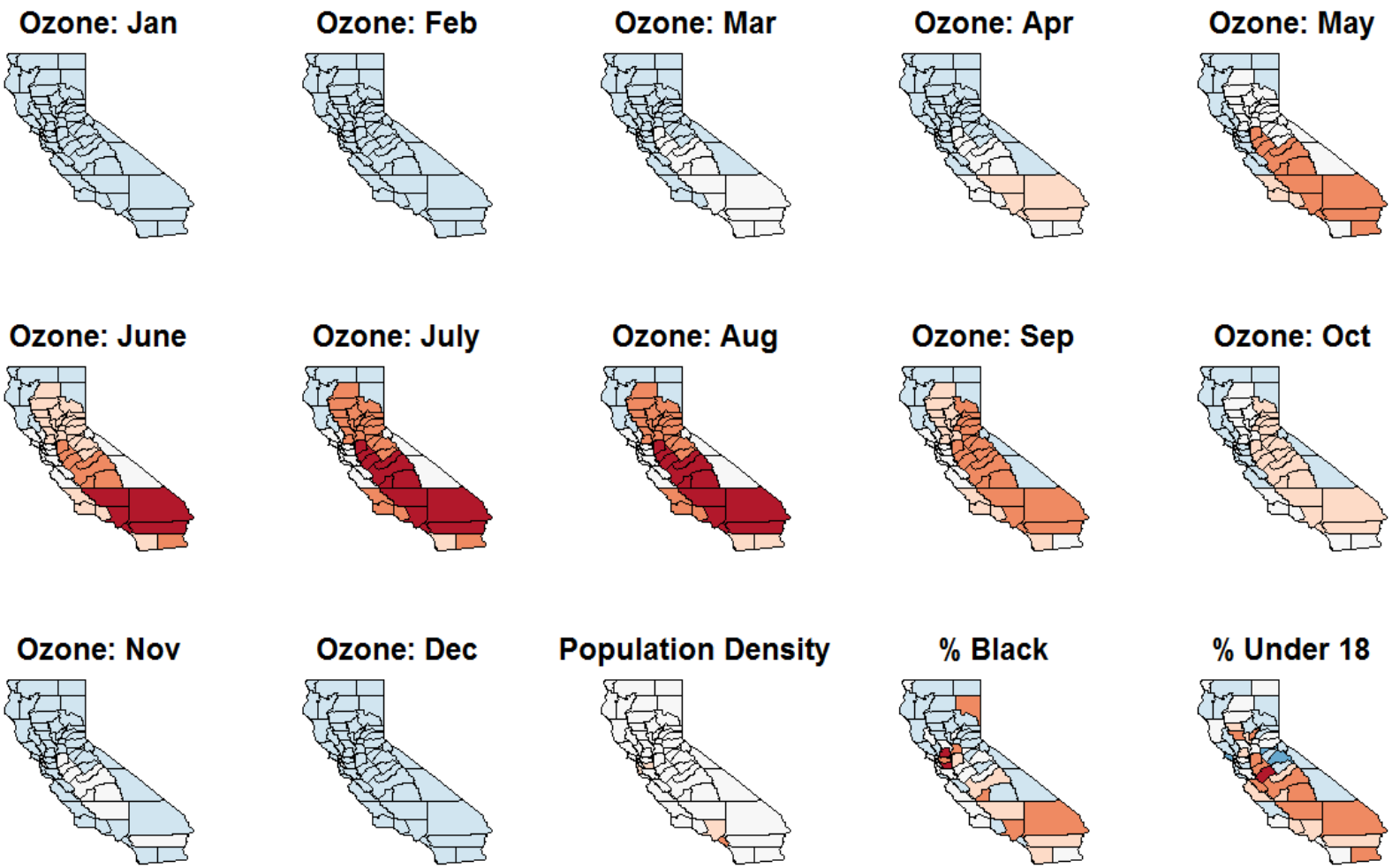
Spatially Misaligned Data



Mapping asthma hospitalizations in California

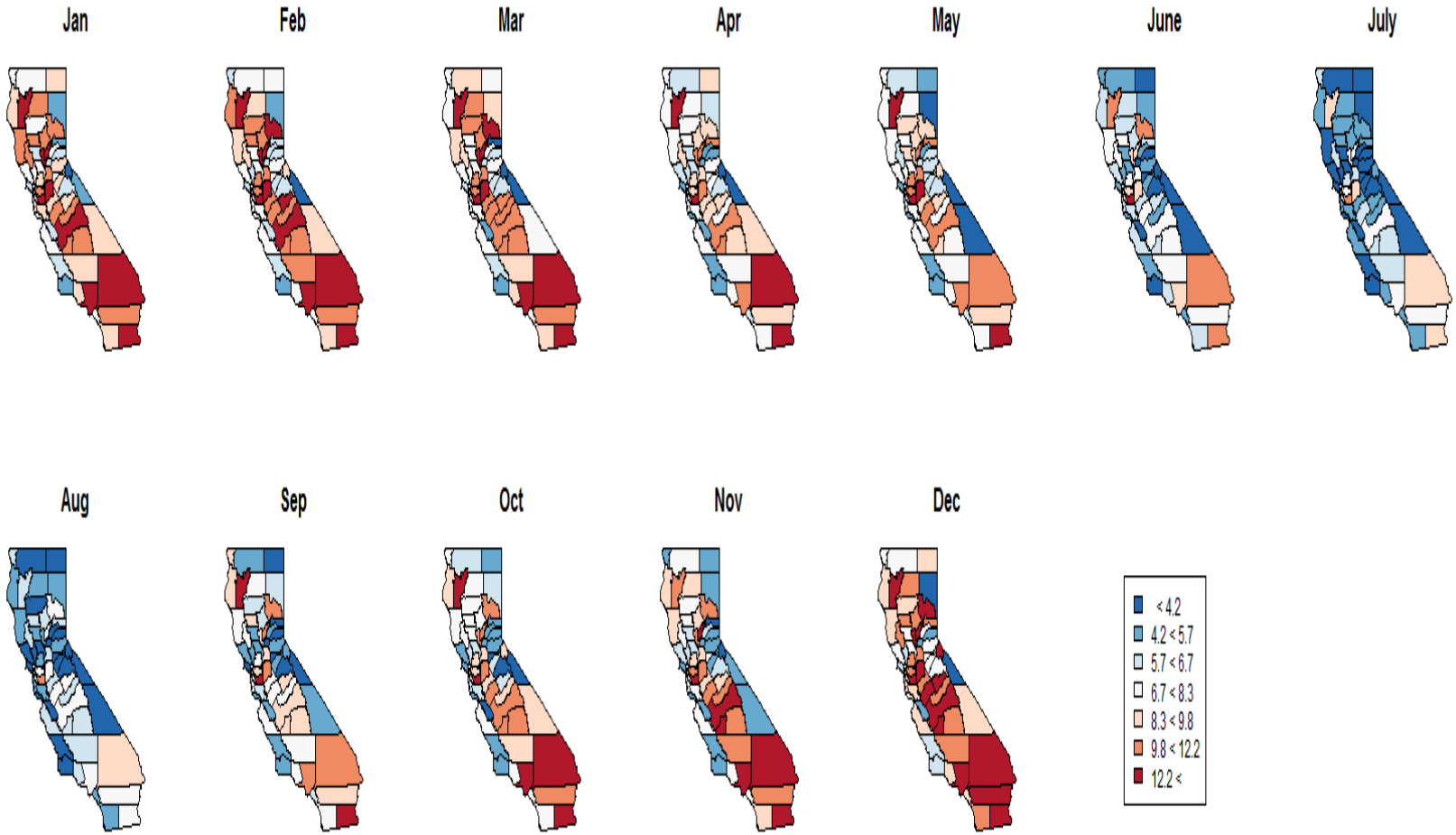


Mapping asthma hospitalizations in California



Mapping asthma hospitalizations in California

$$[\text{Asthma}]_t = [\text{Trends}]_t + [\text{Explanatory}]_t + [(\text{Space-time}) \text{ Interactions}]_t$$



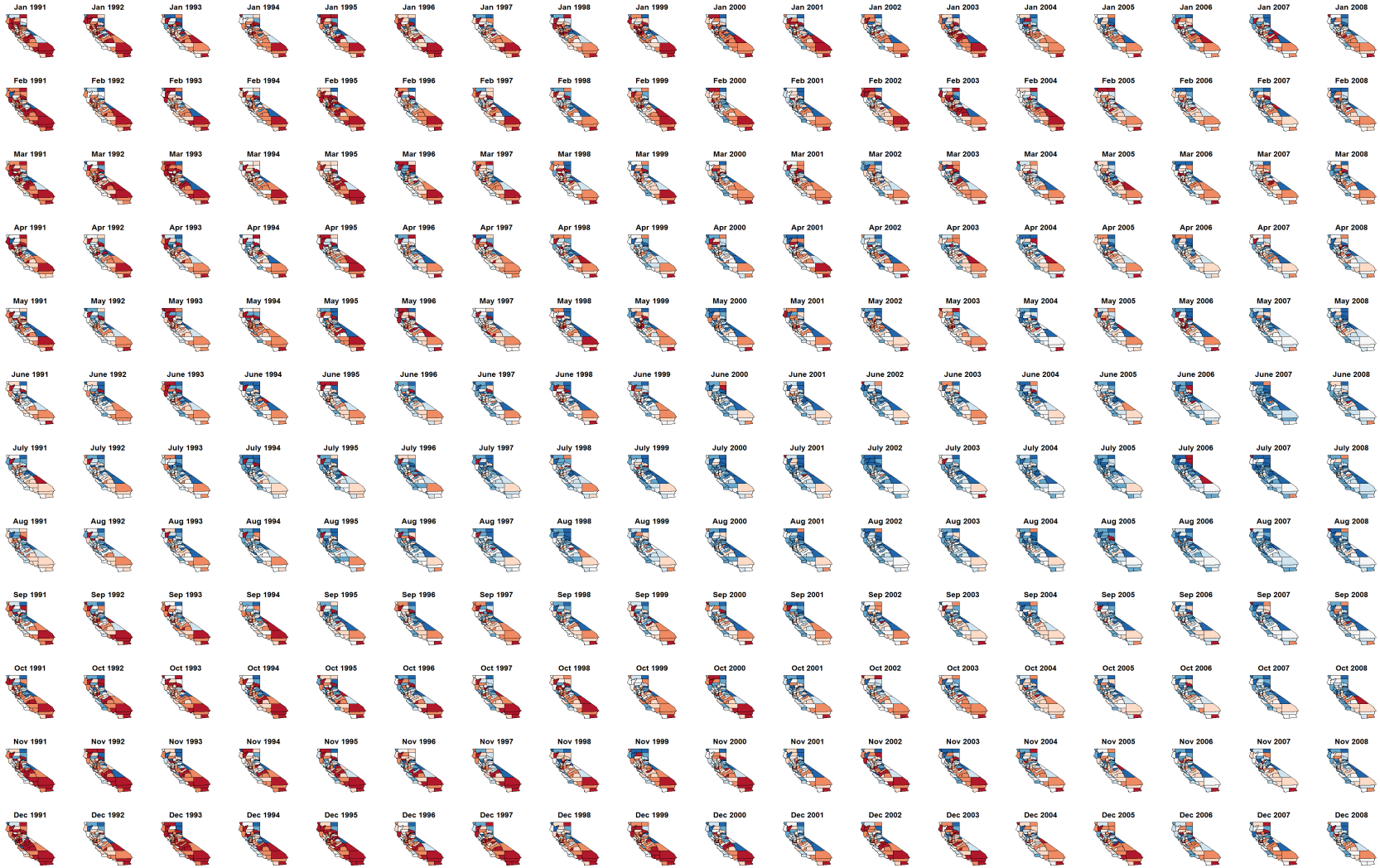
Data Analysis—Parameter Estimates

Parameter	Median (95% CI)	Parameter	Median (95% CI)
β_0 (Intercept)	9.51 (8.46, 10.51)	$\beta_{15} - \beta_{26}$ (Ozone)	
β_1 (Pop Den)	0.63 (0.55, 0.71)	— January	0.51 (-0.95, 1.94)
β_2 (% Black)	1.23 (1.13, 1.33)	— February	0.39 (-0.61, 1.53)
β_3 (% $<$ 18)	1.24 (1.13, 1.34)	— March	0.42 (-0.05, 0.89)
β_4 (Feb)	-0.36 (-1.49, 0.85)	— April	0.21 (-0.05, 0.49)
β_5 (Mar)	-0.24 (-1.32, 0.83)	— May	-0.17 (-0.33, 0.00)
β_6 (Apr)	-1.60 (-2.66, -0.51)	— June	-0.36 (-0.53, -0.20)
β_7 (May)	-1.39 (-2.46, -0.30)	— July	-0.22 (-0.35, -0.09)
β_8 (June)	-2.46 (-3.59, -1.37)	— August	-0.20 (-0.33, -0.07)
β_9 (July)	-3.29 (-4.47, -2.19)	— September	-0.28 (-0.42, -0.12)
β_{10} (Aug)	-3.16 (-4.33, -2.08)	— October	0.06 (-0.13, 0.25)
β_{11} (Sep)	-1.94 (-3.03, -0.88)	— November	0.52 (0.03, 1.05)
β_{12} (Oct)	-1.78 (-2.82, -0.70)	— December	3.15 (1.43, 5.08)
β_{13} (Nov)	-0.87 (-1.94, 0.24)	Spatial smoothing	0.88 (0.85, 0.90)
β_{14} (Dec)	2.42 (1.12, 3.64)	Temporal decay	1.24 (1.18, 1.30)

Table: Estimates and Credible Values of variables on asthma hospitalization rates.

Spatial-Temporal BIG DATA Analytics

$$[\text{Asthma}]_t = [\text{Intercept}]_t + [\text{Explanatory Variables}]_t + [\text{Space-time Interactions}]_t$$



Example 1: U.S. forest biomass data

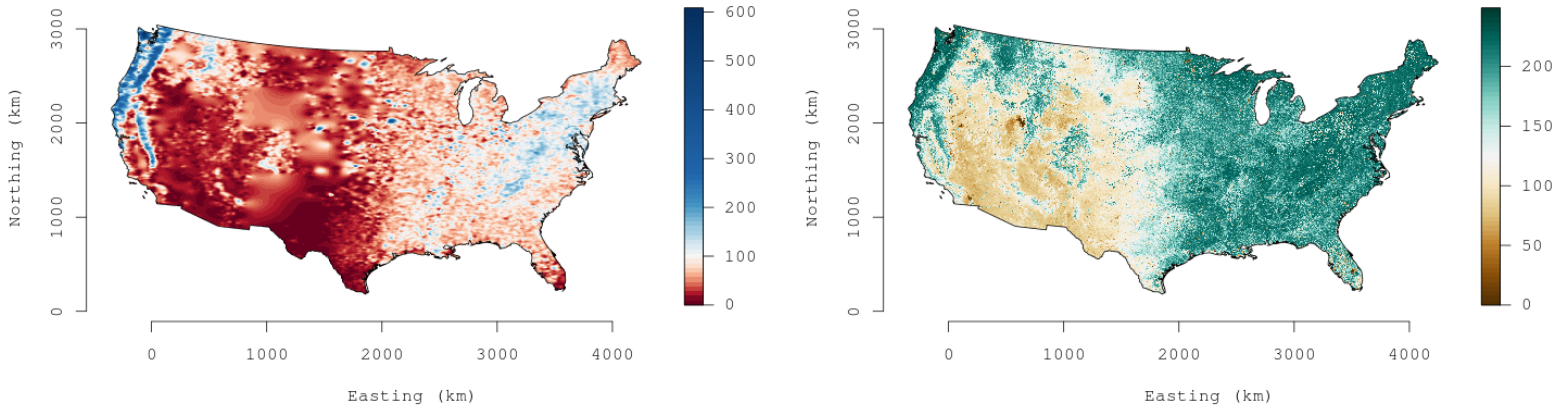
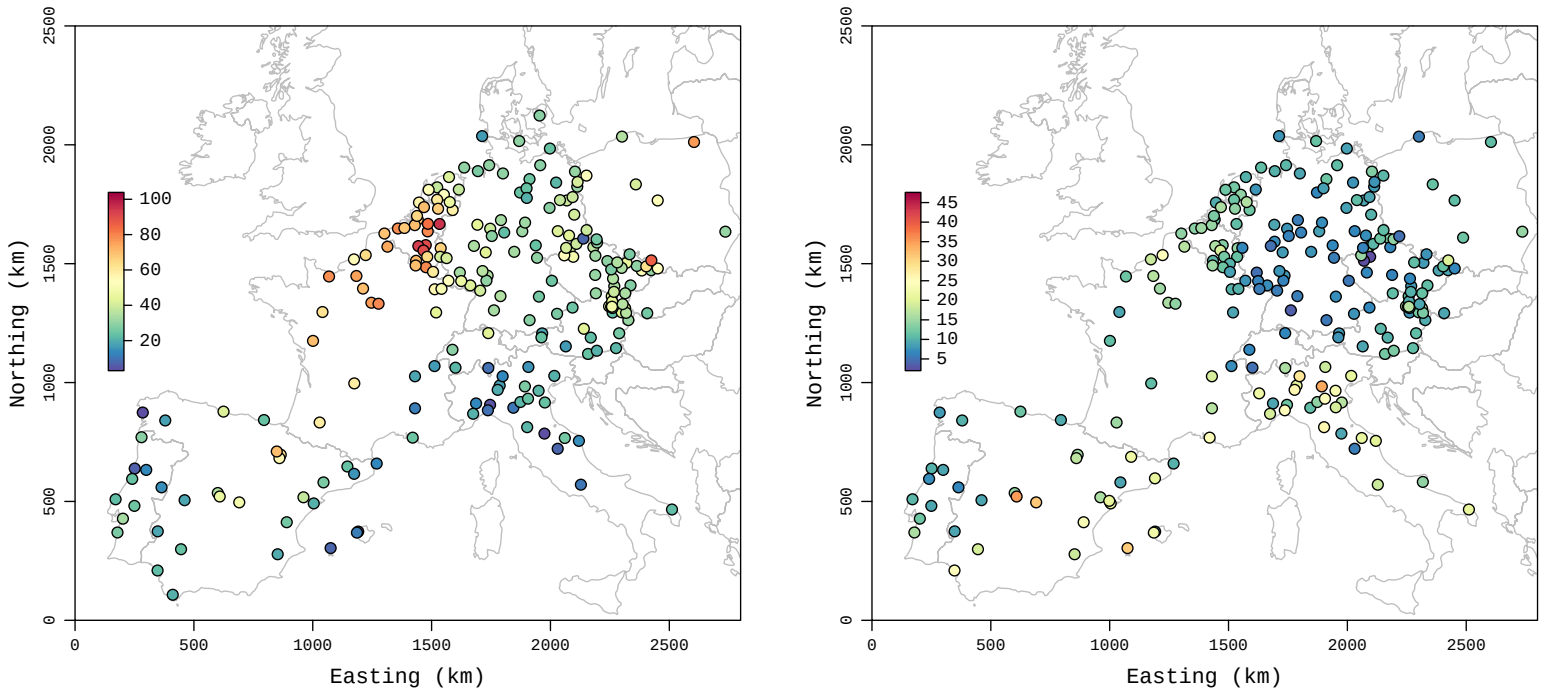


Figure: Observed biomass (left) and NDVI (right)

- Forest biomass data collected over 114,371 plots
- Normalized Difference Vegetation Index (NDVI) is a measure of greenness
- Forest Biomass Regression Model:
  $Biomass(\ell) = \beta_0(\ell) + \beta_1(\ell) NDVI(\ell) + error$

Spatial-Temporal BIG DATA Analytics

Example 2: European Particulate Matter (PM_{10}) data

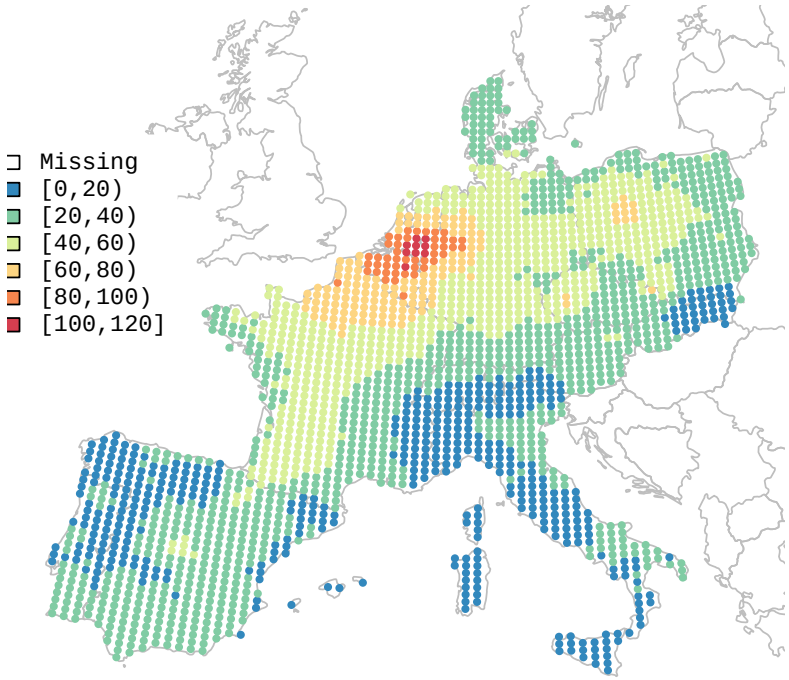


(a) PM_{10} levels in March, 2009 (b) PM_{10} levels in June, 2009

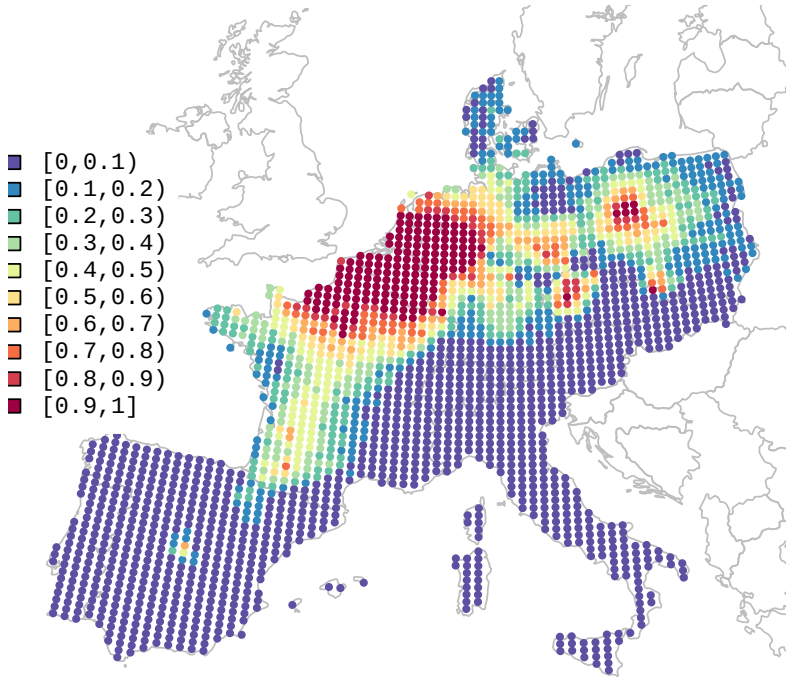
- Significant variation across space and time
- Daily observations at 308 stations for 2 years i.e., $n = 308 \times 730 = 224,840$

Statistical mapping with BIG DATA

European PM_{10} Dataset



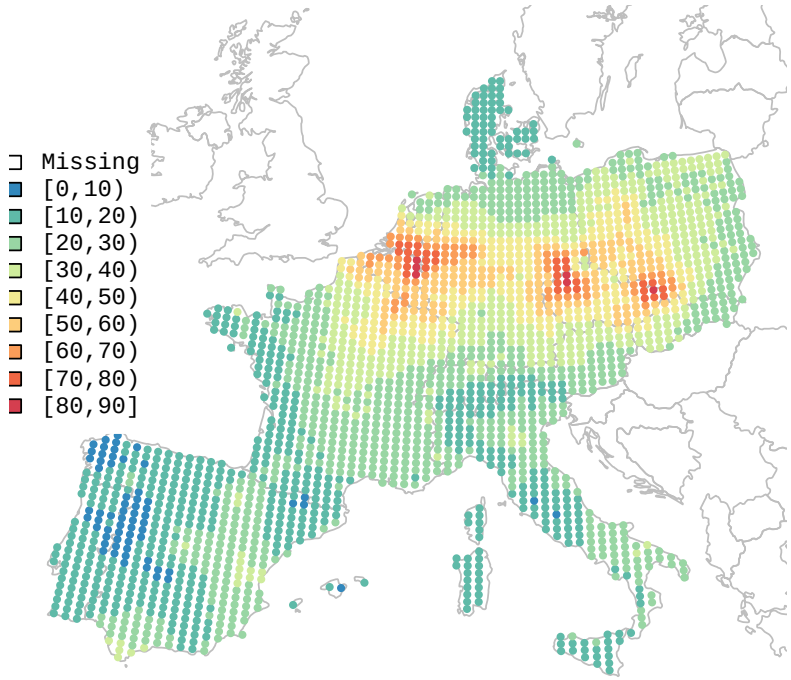
(c) $\widehat{\text{PM}_{10}}$ for 04.03.2009



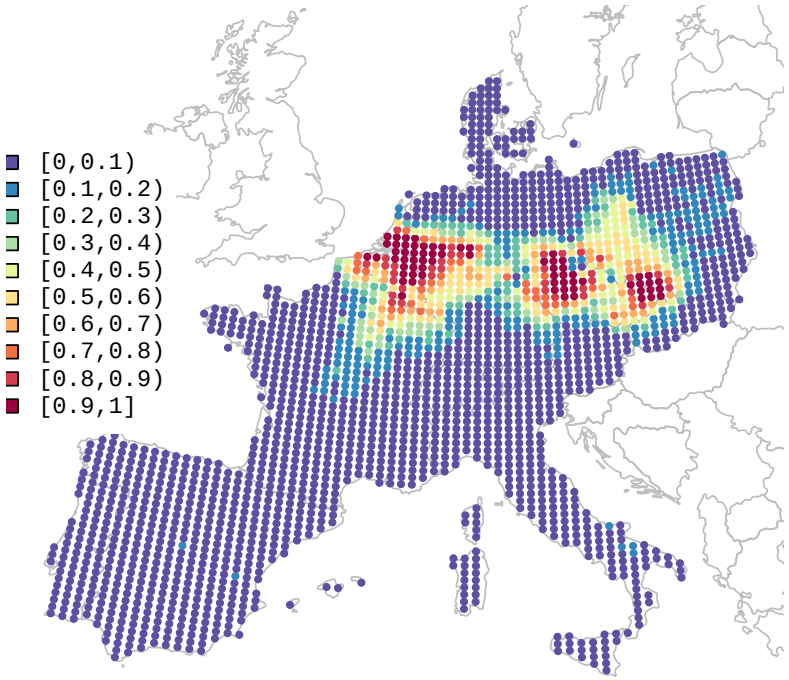
(d) $Pr(\widehat{\text{PM}_{10}} > 50\mu gm^{-3})$

Statistical mapping with BIG DATA

European PM_{10} Dataset

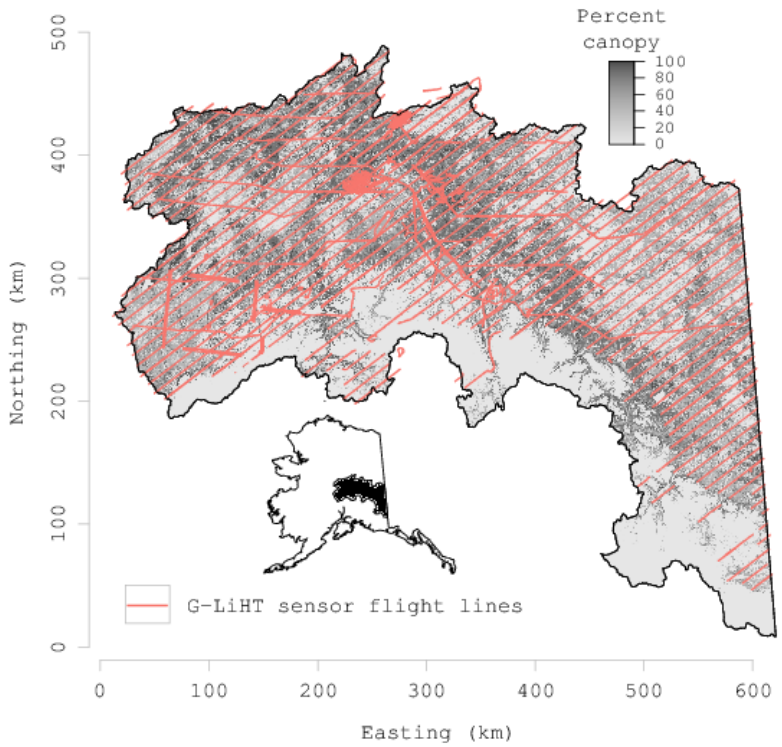


(a) $\widehat{\text{PM}_{10}}$ for 04.05.2009

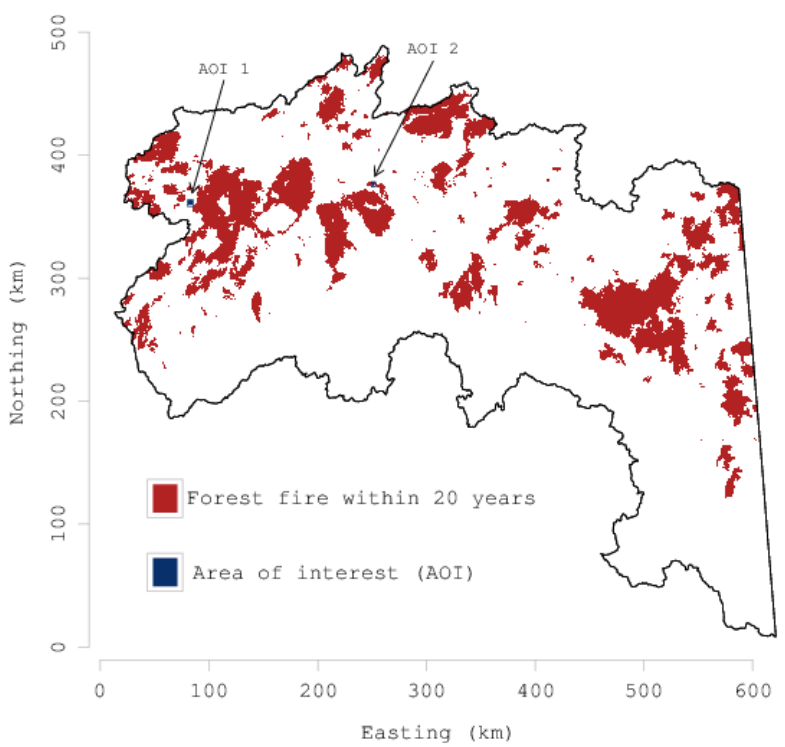


(b) $Pr(\widehat{\text{PM}_{10}} > 50\mu gm^{-3})$

Case Study: Tanana Valley Forest Height Analysis



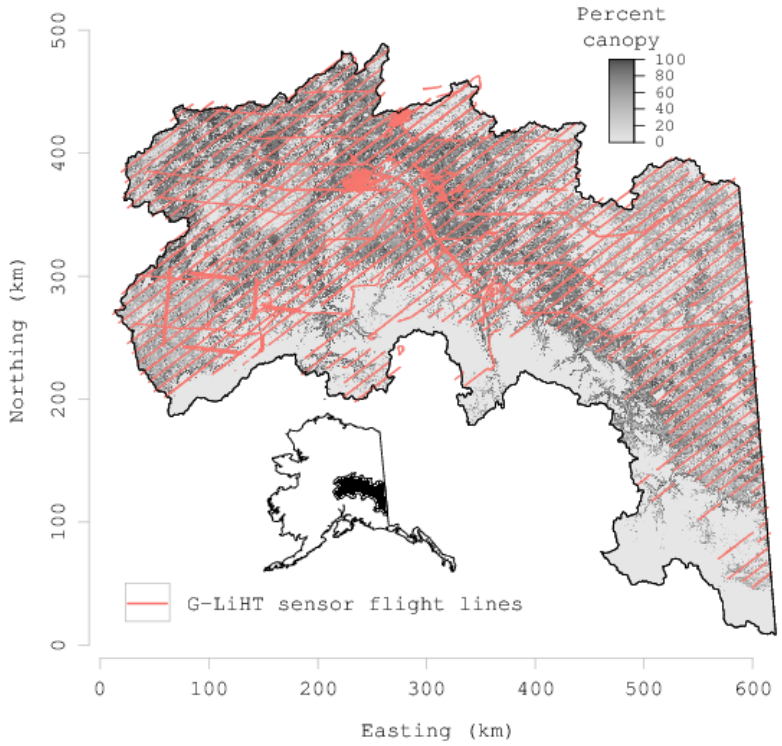
(c) Forest height and tree cover



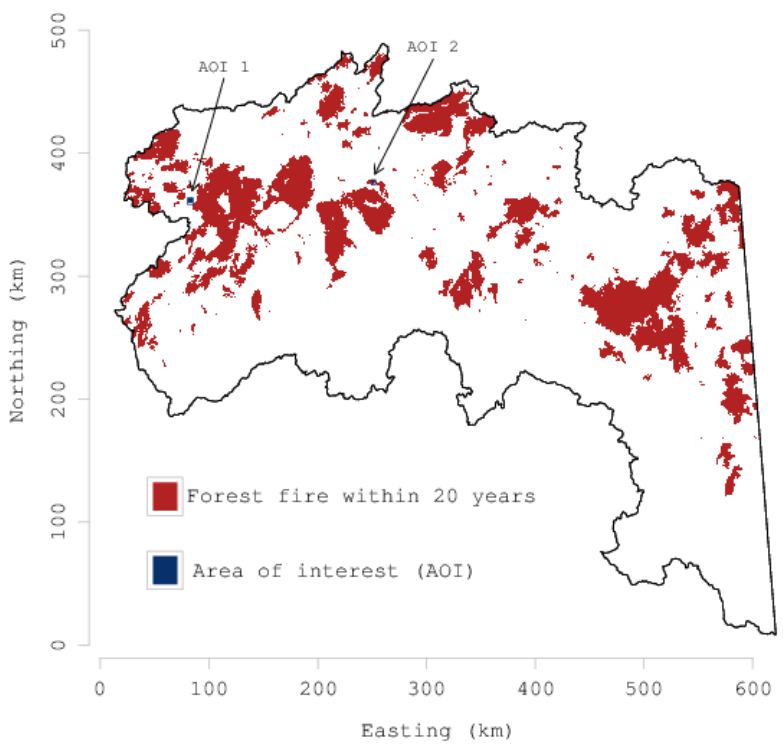
(d) Forest fire history

- Forest height (red lines) data from LiDAR at 40×10^6 locations
- Knowledge of forest height is important for biomass assessment, carbon management etc

Case Study: Tanana Valley Forest Height Analysis



(e) Forest height and tree cover



(f) Forest fire history

- Goal: High-resolution domainwide prediction maps of forest height
- Biomass $=$ [Mean] $+$ [Tree Cover] $+$ [Forest Fire] $+$ [Spatial Process]
- Methodology: Entire model-based spatial data analysis done is 2 minutes on a modest laptop.

Selected papers on Spatial BIG DATA analysis

- Banerjee, S. (2017). High-dimensional Bayesian geostatistics. Bayesian Analysis, 12, 583–614. DOI: http://dx.doi.org/10.1214/17-BA1056R
- Banerjee, S. (2020). Modeling massive spatial datasets using a conjugate Bayesian linear modeling framework. Spatial Statistics, 37, 10047. DOI: https://doi.org/10.1016/j.spasta.2020.100417
- Banerjee, S., Gelfand, A.E., Finley, A.O. and Sang, H. (2008). Gaussian predictive process models for large spatial datasets. Journal of the Royal Statistical Society Series B, 70, 825–848. DOI: http://dx.doi.org/10.1111/j.1467-9868.2008.00663.x
- Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process models for large geostatistical datasets. Journal of the American Statistical Association, 111, 800–812. DOI: http://dx.doi.org/10.1080/01621459.2015.1044091
- Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A. S., and Schaap, M. (2016). Non-separable dynamic Nearest-Neighbor Gaussian Process models for large spatio-temporal data with an application to particulate matter analysis. Annals of Applied Statistics, 10, 1286–1316. DOI: http://dx.doi.org/10.1214/16-AOAS931
- Finley, A. O., Datta, A., Cook, B. C., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). Applying Nearest Neighbor Gaussian Processes to massive spatial data sets: Forest canopy height prediction across Tanana Valley Alaska. Journal of Computational and Graphical Statistics. DOI: https://doi.org/10.1080/10618600.2018.1537924
- Guinness, J. (2018). Permutation methods for sharpening Gaussian Process approximations. Technometrics, 60, 415–429.
- Heaton, M., Datta, A., Finley, A., Furrer, R., Guhaniyogi, R., Gerber, F., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D., and Zammit-Mangion, A. (2017). Methods for analyzing large spatial data: A review and comparison. arXiv:1710.05013.
- Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. Journal of the American Statistical Association, 112, 201–214.
- Katzfuss, M. and Guinness, J. (2017). A general framework for vecchia approximations of gaussian processes. arXiv preprint arXiv:1708.06302.
- Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution gaussian process model for the analysis of large spatial datasets. Journal of Computational and Graphical Statistics, 24, 579–599.
- Peruzzi, M., Banerjee, S. and Finley, A.O. (in press). Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains. Journal of the American Statistical Association, DOI: https://doi.org/10.1080/01621459.2020.1833889.
- Taylor-Rodriguez, D., Finley, A.O., Datta, A., Babcock, C., Andersen, H.E., Cook, B.C., Morton, D.C. and Banerjee, S. (2019). Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping. Statistica Sinica, 29, 1155–1180. DOI: https://doi.org/10.5705/ss.202018.0005.
- Zhang, L., Datta, A. and Banerjee, S. (2019). Practical Bayesian modeling and inference for massive spatial datasets on modest computing environments. Statistical Analysis and Data Mining: The ASA Data Science Journal, 12, 197–209. DOI: https://doi.org/10.1002/sam.11413

Thank You!