Humans as Sensors: Citizen Science Data to Assess Climate Change

Prof. Chris Brunsdon

People, Space and Place Research Cluster
School of Environmental Sciences
Roxby Building
The University of Liverpool
Liverpool L69 7ZT

August 2012

Geoinformatica
DOI 10.1007/s10707-012-0159-6

Assessing the changing flowering date of the common lilac in North America: a random coefficient model approach

Chris Brunsdon · Lex Comber

Received: 5 October 2011 / Revised: 1 May 2012 /
Accepted: 14 June 2012

Abstract A data set consisting of *Volunteered geographical information* (VGI) and data provided by expert researchers monitoring the first bloom dates of lilacs from 1956 to 2003 is used to investigate changes in the onset of the North American spring. It is argued that care must be taken when analysing data of this kind, with particular focus on the issues of lack of experimental design, and Simpson's paradox. Approaches used to overcome this issue make use of random coefficient modelling, and bootstrapping approaches. Once the suggested methods have been employed, a gradual advance in the onset of spring is suggested by the results of the analysis. A key lesson learned is that the appropriateness of the model calibration technique used given the process of data collection needs careful consideration.

Keywords Phenology · Random effects models · Citizen science

Overview

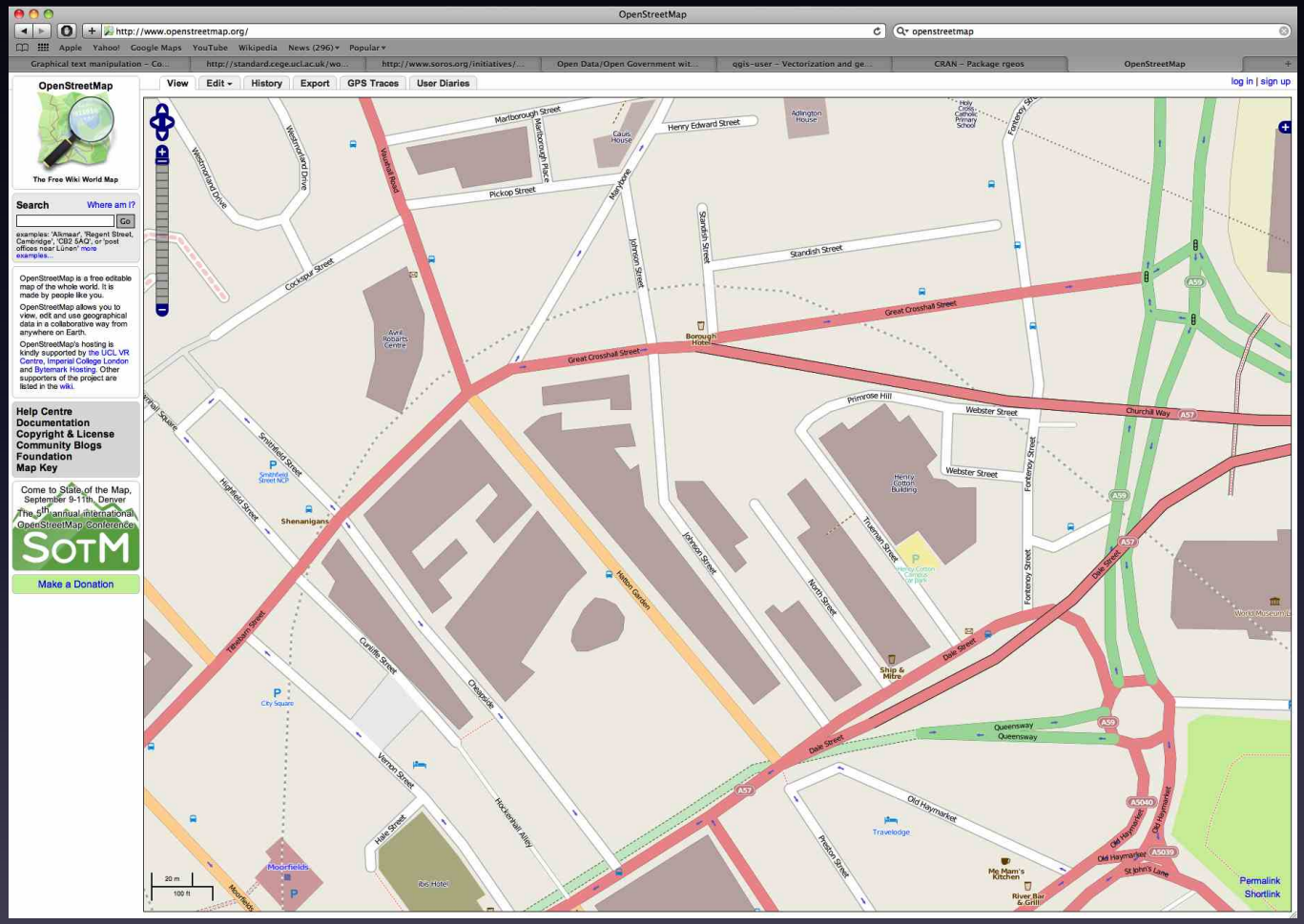
- Three Ideas About Data
  - Open Data
  - Volunteered Geographical Information
  - Citizen Science
- Methodological Issues
  - Implications for Data Analysis
- Illustration by Example
  - Monitoring Climate Change via Citizen Science

Volunteered Geographic Information

- See for example Goodchild in GeoJournal (2007)
  - DOI: 10.1007/s10708-007-9111-y

'... the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies.'

Example:Open Streetmap



Is VGI a new outlet for Citizen Science?

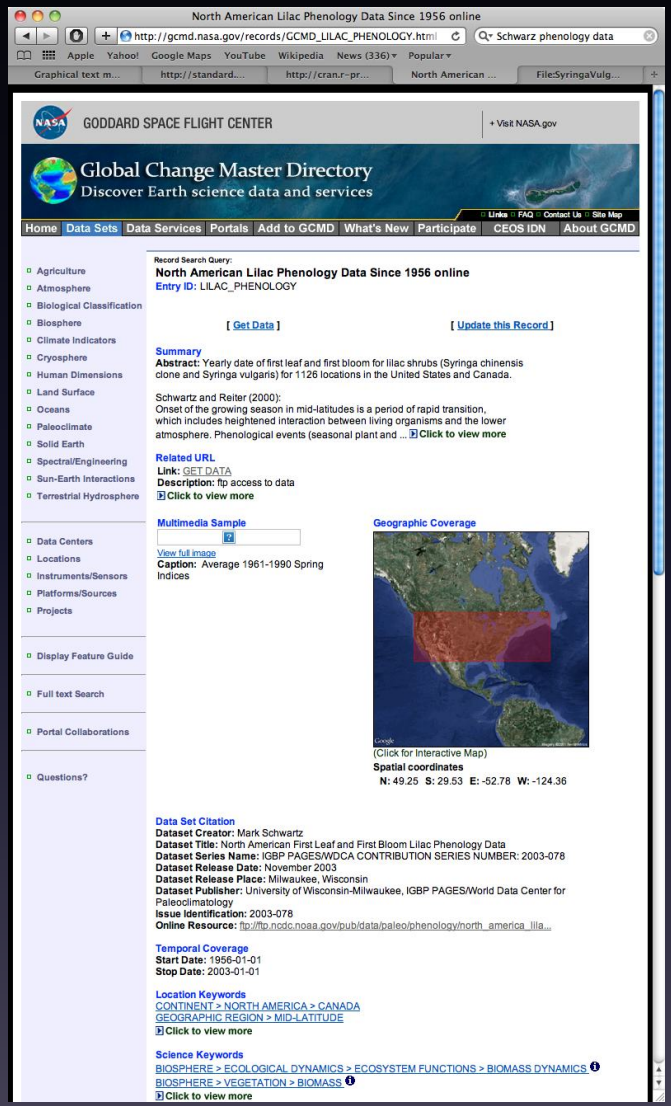
Issues analysing Open Data

- Quite a lot of this data is geotagged and time-stamped
  $\Rightarrow$ in some way
- But there are special special issues
  - **Experimental design**
    $\Rightarrow$ there often is none!
  - **Metadata**
    $\Rightarrow$ Not always formal...

Climate Change: First Bloom of North American Lilac

- Monitoring first bloom dates of certain plants
- Warmer climate $\Rightarrow$ earlier dates ?
- Here we consider the lilac in North America

Open data source (VGI):



Source created by Mark Schwartz

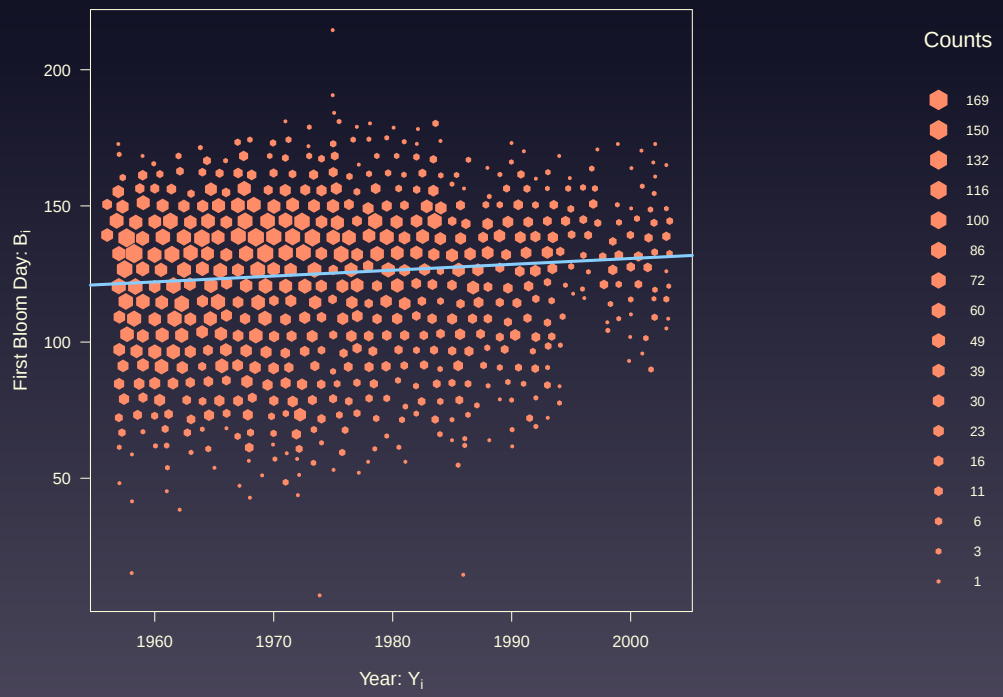
Details

- Data consists of:
- First bloom dates for years **1956-2003**
- at **1126** observation locations
- across USA (with a handful of observations just outside)
- A total of **14265** observations
- **Latitude and longitude** of each observation location is also recorded

Results

$d_{ij} = S + Ry_{ij} + \epsilon_{ij}$:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0000	0.1861	-0.00	1.0000
year	0.2150	0.0178	12.11	0.0000



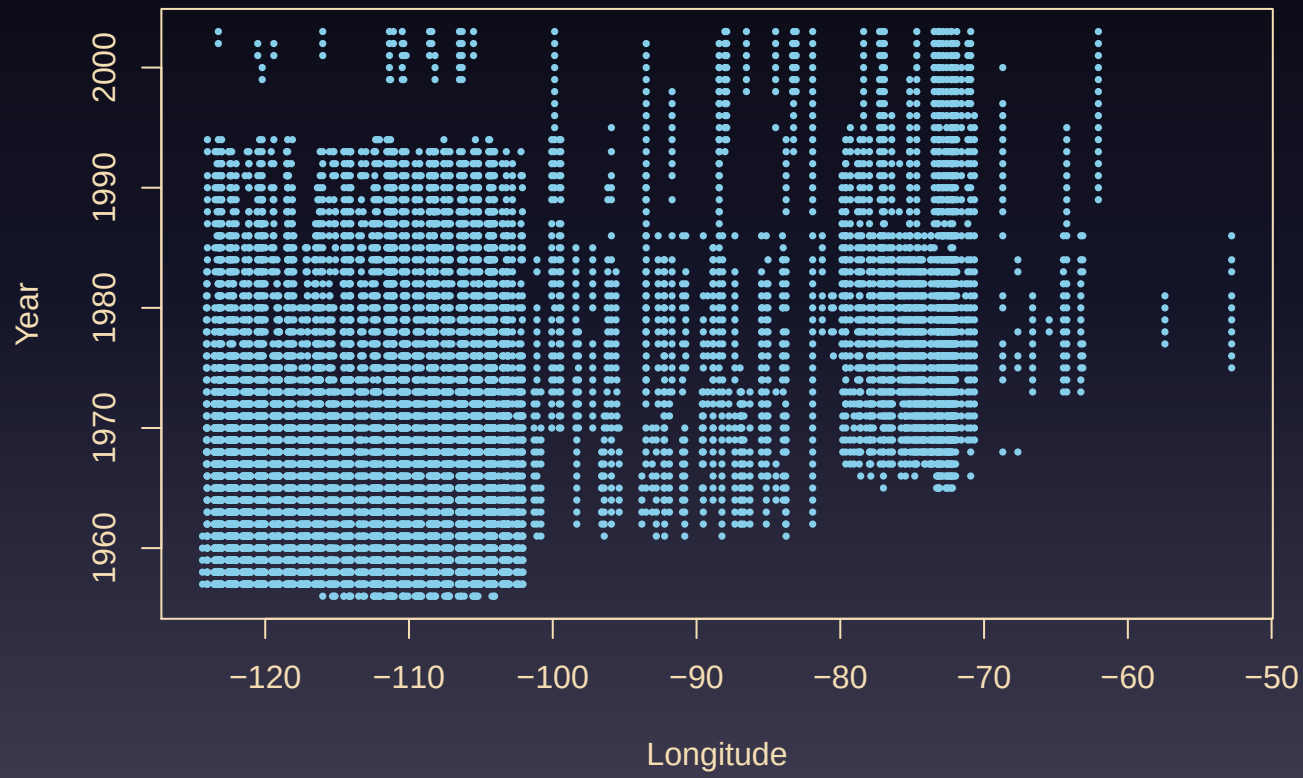
What !?!

- A rather surprising result
- Implies first bloom date getting later by about 1 day every 5 years
- **opposite of expected effect!**

Consider things further:

- How evenly balanced are the observations?
- What **geographical** effects might influence first bloom dates?
  - Elevation
  - Onset of spring
  - Average annual temperature
  - Soil type

Even space/time balance?



In tabular form

	Eastern	Western
1955-1959	0	1997
1960-1964	97	2548
1965-1969	449	2420
1970-1974	640	1965
1975-1979	643	1049
1980-1984	578	778
1985-1989	248	664
1990-1994	231	411
1995-1999	183	8
2000-2004	120	43

Split the Data into Two

	Estimate	Std. Error	t value	Pr(>|t|)
a (Eastern)	134.017	0.406	330.473	0.000
a (Western)	122.382	0.315	388.282	0.000
b_Y (Eastern)	0.048	0.041	1.164	0.244
b_Y (Western)	0.033	0.021	1.555	0.120

Table: Regression analysis for models fitted separately to each network

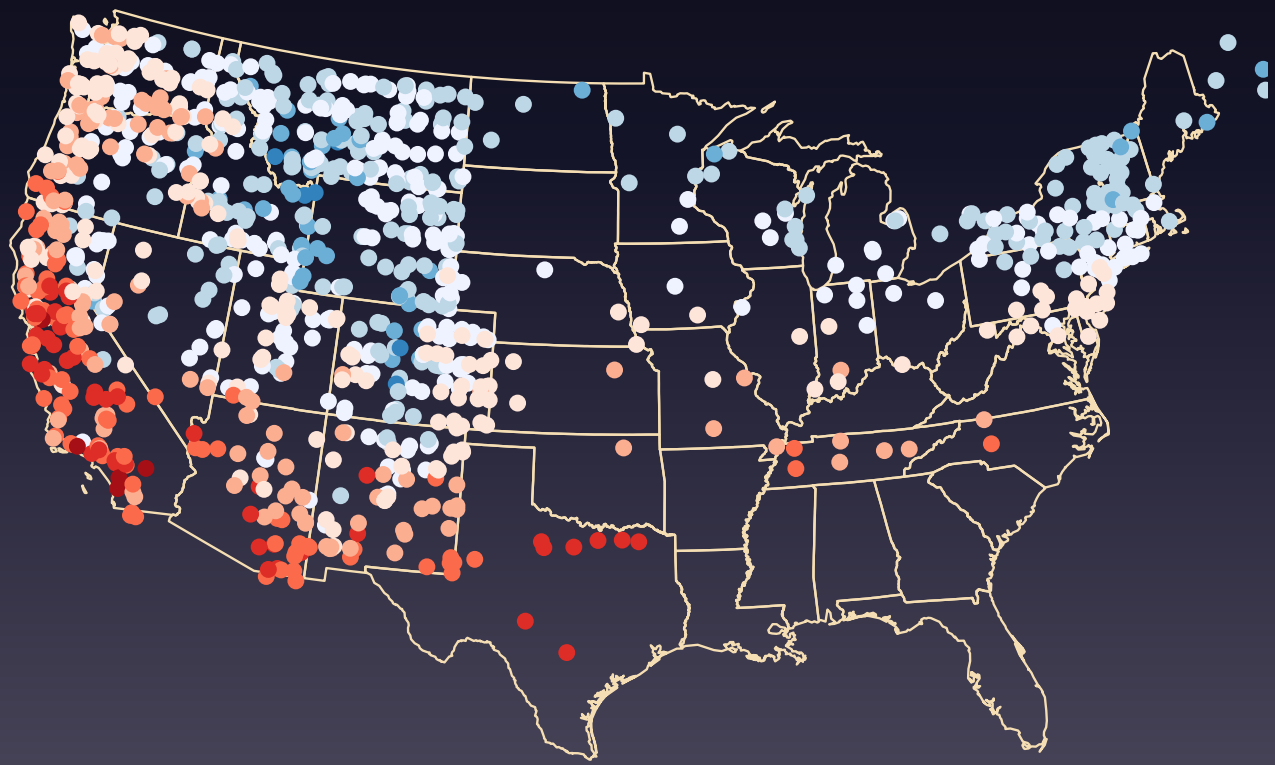
An Explanation and A New Approach

- The 'green wave'
  - Spring comes sooner in some parts of the US
- In the regression model, allow intercept to change between locations
- This allows for the 'different baseline' problem.
- The random coefficient model:
- Intercept in regression is a random variable at the locational level

Estimates of Random Coefficients

Random Coefficient Model (MLM)

$$d_{ij} = s_i + Ry_{ij} + \epsilon_{ij}$$



Results

$d_{ij} = s_i + Ry_{ij} + \epsilon_{ij}$:

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-0.95	0.64	13139.00	-1.50	0.13
year	-0.18	0.01	13139.00	-17.18	0.00

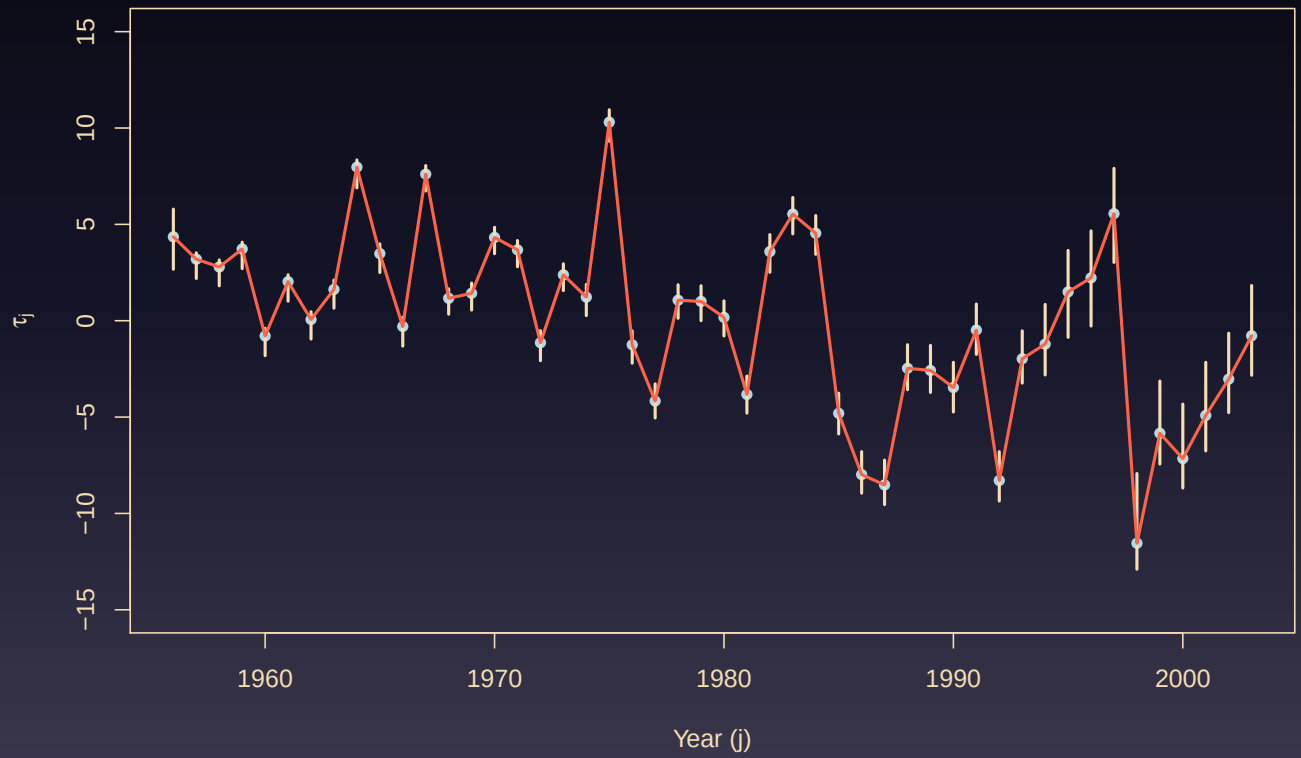
- Note the change in sign of the **year** coefficient
- Implies first bloom date getting earlier by about **1 day every 6 years**
- This estimate allows for **imbalance** in observation locations over time

A Non-Linear Time Effect

$$d_{ij} = A + \tau_i + \upsilon_j + \epsilon_{ij}$$

- τ_i Time effect
- υ_j Observation station effect
- Thus, variation over time is modelled as an individual random variable each year
- ... in the same way as geographical effect
- ϵ_{ij} is a Gaussian error term

Results

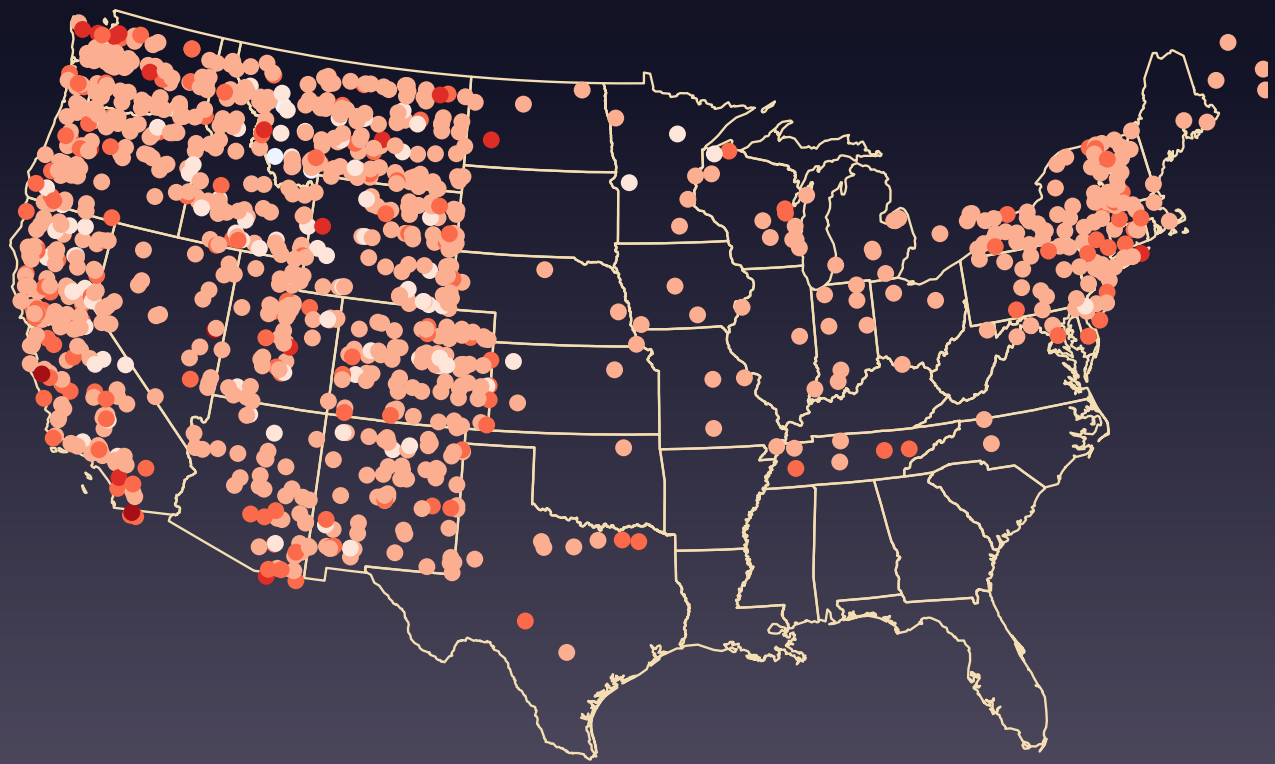


Estimates of Random Coefficients

Random Coefficient Model (MLM) - Varying Slope

$$d_{ij} = s_i + R_i y_{ij} + \epsilon_{ij}$$

PS. Why not GWR?



Lessons Learned

- The key thing here is **working without experimental design**.
- The lack of design can lead to misleading outcomes
  $\Rightarrow$ Unless you use an appropriate technique
- Learning **from** data is good
  $\Rightarrow$ You need to learn **about** the data as well.
- Statistical analysis is an **aid** to intelligent thought, not a **substitute** for it!
- Need to think about **space** and **geography**...

Thank You

References

Schwartz, M.D. and Reiter, B.E. **Changes in North American Spring**
International journal of Climatology,2000,v20,pp929–932

Laird, N.M. and Ware, J.H. **Random-Effects Models for Longitudinal Data**
Biometrics,1982, v38, pp963–974

Brunsdon, C. and Comber, A. **Assessing the changing flowering date of the common lilac in North America: a random coefficient model approach**
GeoInformatica: DOI:10.1007/s10707-012-0159-6

Multilevel Models in R
http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf

Random Effects in R
http://www.stat.wisc.edu/courses/st572-larget/Spring2007/handouts13-2.pdf