Interactive Visualization for Missing Values, Time Series, and Areal Data

PhD Thesis Proposal

by

Xiaoyue Cheng

Program of Study Committee:

Dianne Cook

Heike Hofmann

Petrutza Caragea

Zhengyuan Zhu

Alan Wanamaker

April 2014

- Education
  - B.A. in Statistics, Renmin University of China, 2002-2006
  - M.A. in Statistics, Renmin University of China, 2006-2009
  - Statistics, University of Kentucky, fall 2009
  - Statistics, Iowa State University, 2010-Present
- Internship
  - Google Summer of Code, summer 2011
  - Amazon.com, summer 2012 & 2013

Motivation

- Visualization is used to explore data, examine variation, reveal trends, and diagnose models.
- Challenge: big data
  - Solution: a large number of pictures
- Challenge: creation, storage, and querying the pictures
  - Solution: interactive graphics

Previous research

- Seminal work: PRIM-9(Fisherkeller et al., 1988), Dataviewer(Hurley, 1987), XLispStat(Tierney, 2009), DataDesk(Velleman, 1989), XGobi(Swayne et al., 1998), MANET(Unwin et al., 1996)
- Contemporary software: Mondrian(Theus, 2003), GGobi(Swayne et al., 2003), and R packages iplots(Urbanek and Theus, 2003), cranvas(Xie et al., 2013), ggvis(RStudio, 2014)
- From information visualization community: Tableau(Hanrahan et al., 2007), d3.js(Bostock, 2012), Processing(Reas and Fry, 2007)

Grammar of graphics

- Static graphics
  - Cleveland and McGill (1987) studied human perception related to data plots;
  - Wilkinson et al. (2006) and Wickham (2009) developed a grammatical construction.
- Interactive graphics
  - Buja et al. (1988) studied the elements of a viewing pipeline;
  - Wickham et al. (2009) discussed the designs for controlling the flow of updates;
  - Lawrence and Verzani (2012) introduced the fundamentals of graphical user interfaces (GUI);
  - Xie et al. (2014) introduced the MVC patterns in interactive graphics.

Problems

- General manipulations are insufficient for specific data types and exploratory purposes
- Preference of the integration between interactive graphics and statistical analysis tools
- Lack of the grammar for the specific manipulations
- Three aspects in the thesis:
  - Missing data
  - Temporal and longtudinal data
  - Spatial (areal) data

Overview of the thesis

- Missing data (Chapter 2)
  - Goal: explore missing value structure, examine missingness assumptions, compare imputation results using static plots and numerical summaries
  - Implementation: GUI
- Temporal and longtudinal data (Chapter 3 & 4)
  - Goal: explore the trends, seasonality or unusual individuals
  - Implementation: interactive graphics
- Areal data (Chapter 5)
  - Cartograms: emphasize the important areas
  - Goal: create and tune the cartogram
  - Implementation: GUI & interactive graphics

Visually Exploring Missing Values in Multivariable Data Using a Graphical User Interface

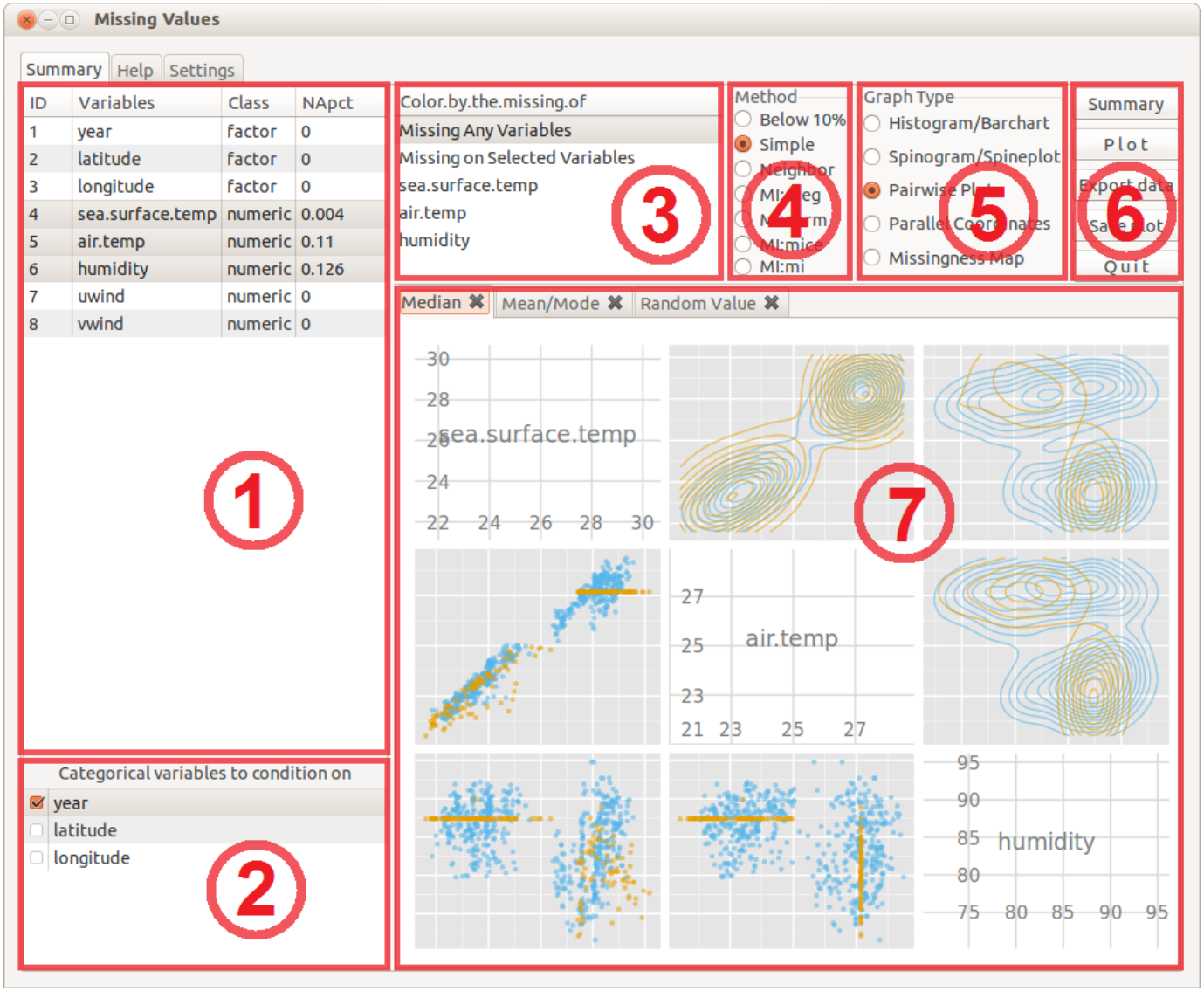
Existing work on visualization

- Independent software
  - MANET(Missings Are Now Equally Treated;Unwin et al. 1996)
  - XGobi and GGobi(Swayne and Buja, 1998)
- From R community
  - VIM(Templ et al., 2013) & VIMGUI(Schopfhauser et al., 2013)
  - mi(Su et al., 2011) & migui(Lee and Su, 2011)
  - Amelia(Honaker et al., 2011) & AmeliaView
  - miP(Brix, 2012) for mice, mi, and Amelia

MissingDataGUI

- A visualization tool to
  - explore missing value structure
  - examine missingness assumptions
  - compare imputation results using static plots and numerical summaries
- Compared to the previous work, it provides
  - multiple choices of imputation and visualization methods for various data types
  - imputation conditioned on factors
  - easy comparison between methods and between chains from multiple imputations
  - assistance to examine the assumption and find the dependent variables
  - user-friendly design

Design


Missing Values
Summary
Help
Settings
ID
Variables
Class
NApct
1 year factor 0
2 latitude factor 0
3 longitude factor 0
4 sea.surface.temp numeric 0.004
5 air.temp numeric 0.11
6 humidity numeric 0.126
7 uwind numeric 0
8 vwind numeric 0
1
Categorical variables to condition on
year
latitude
longitude
2
Color.by.the.missing.of
Missing Any Variables
Missing on Selected Variables
sea.surface.temp
air.temp
humidity
3
Method
Below 10%
Simple
Neighbor
MI:reg
MI:rm
MI:mice
MI:mi
4
Graph Type
Histogram/Barchart
Spinogram/Spineplot
Pairwise Pl
Parallel Coordinates
Missingness Map
5
Summary
Plot
Export data
Save plot
Quit
6
Median
Mean/Mode
Random Value
sea.surface.temp
air.temp
humidity
7

Summary of missing values

Numeric Summary for Missing Values

year=1997

Missing:
2.62% of the numbers
12.5% of variables
20.92% of samples

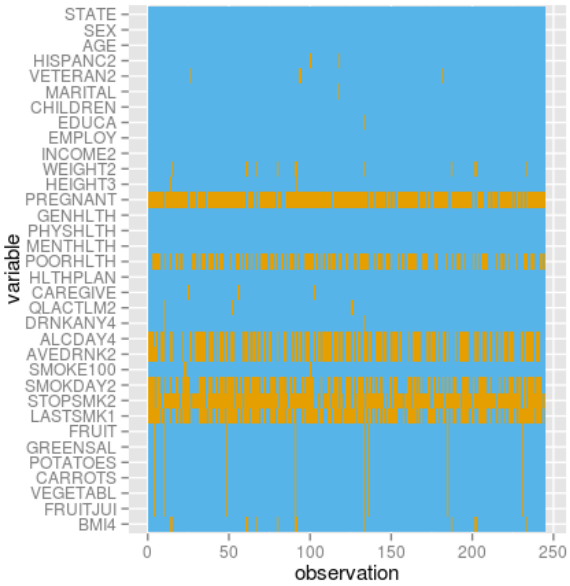
No_of_miss_by_case	No_of_Case	Percent
0	291	79.1
1	77	20.9
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0

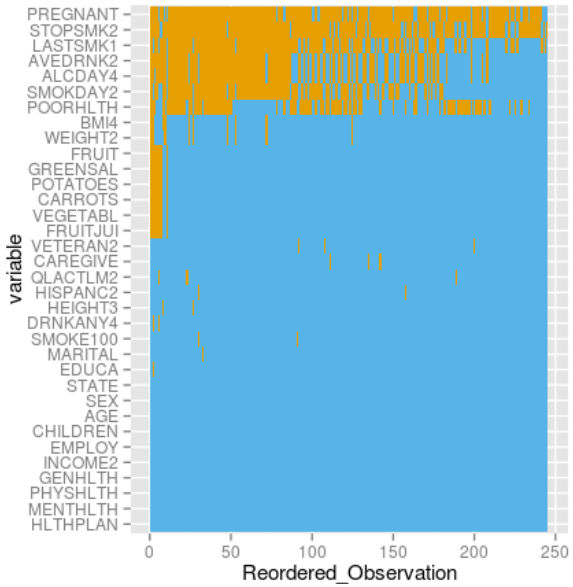
year=1993

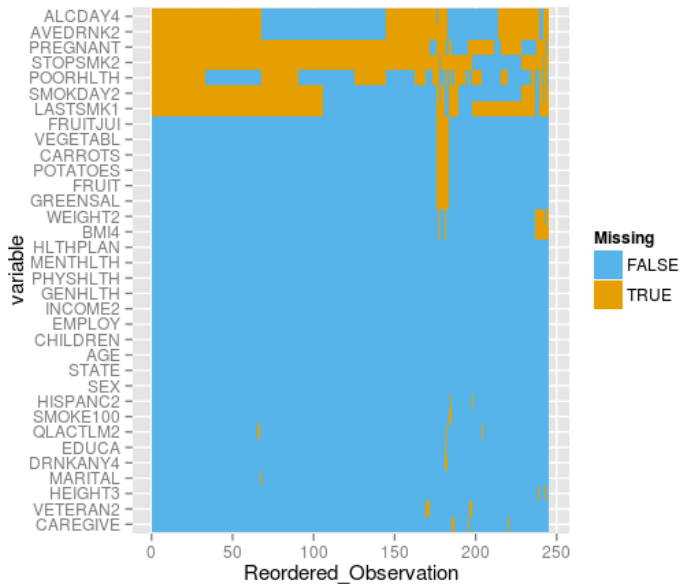
Missing:
3.4% of the numbers
37.5% of variables
25.54% of samples

No_of_miss_by_case	No_of_Case	Percent
0	274	74.5
1	90	24.5
2	2	0.5
3	2	0.5
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0

Missingness map





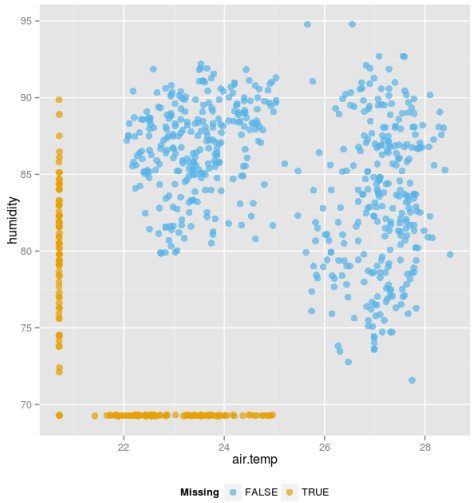


Methods

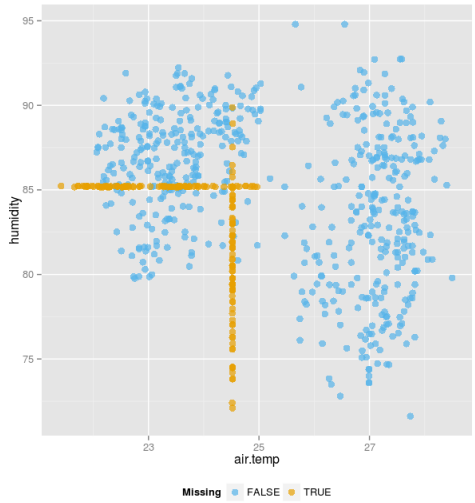
Method	Description	Determinisitic	Univariate	Multiple imp.
Below 10%	below 10% of the range	x	x	
Simple	overall median	x	x	
	overall mean/mode	x	x	
	random value		x	
Neighbor	mean of the nearest neighbors	x		
	random nearest neighbor			
MI:areg	predictive mean matching			x
MI:norm	multivariate normal model			x
MI:mice	multivariate imp. by chained equations			x
MI:mi	multiple iterative regression imputation			x

Univariate imputations

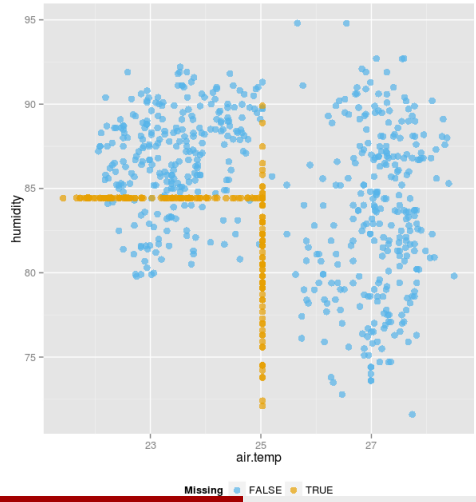
(a) Below 10%



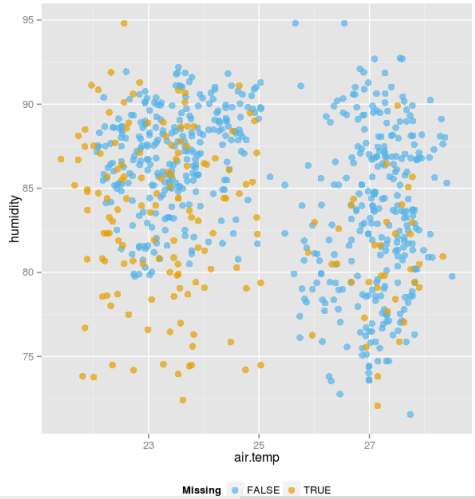
(b) Overall median



(c) Overall mean

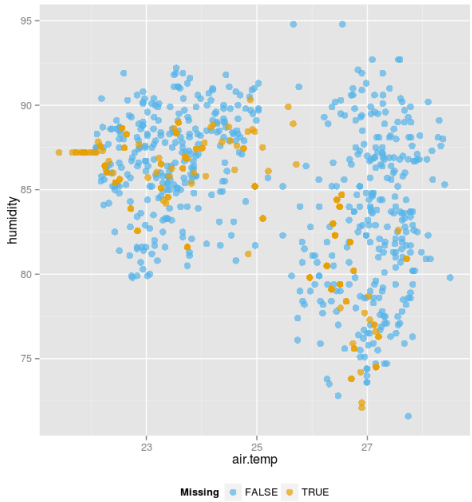


(d) Random value

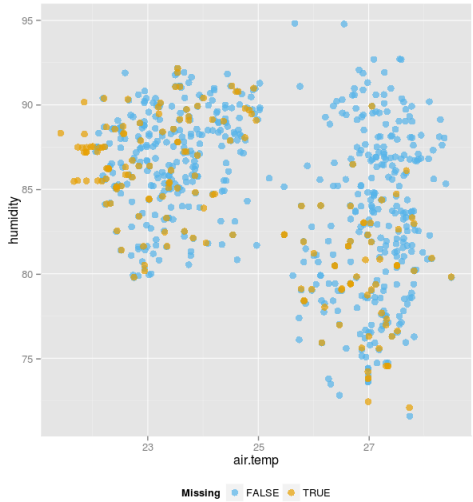


Nearest neighbor imputations

(a) Mean of the neighbors



(b) Random neighbor



Select variables → **Get the data** → **Loop for each incomplete case** → **Combine with the complete cases** → **Standardize the complete columns** → **Compute the distance and rank neighbors**

ID	Variables	Class	NApct
1	year	factor	0
2	latitude	factor	0
3	longitude	factor	0
4	sea.surface.temp	numeric	0.004
5	air.temp	numeric	0.11
6	humidity	numeric	0.126
7	uwind	numeric	0
8	vwind	numeric	0

air.temp	humidity	uwind	vwind
23.54	92.2	-2.5	0.9
23.43	89.9	-2.7	1.1
23.89	88.6	-2.1	0.4
23.79	91.6	-2.3	1.5
23.19	NA	-5.3	1.8
23.50	NA	-4.7	3.1
22.87	NA	-7.1	2.5
NA	NA	-3.0	1.9
NA	NA	-5.6	3.1

air.temp	humidity	uwind	vwind
23.19	NA	-5.3	1.8

air.temp	humidity	uwind	vwind
23.54	92.2	-2.5	0.9
23.43	89.9	-2.7	1.1
23.89	88.6	-2.1	0.4
23.79	91.6	-2.3	1.5
23.19	NA	-5.3	1.8

air.temp	uwind	vwind
-0.09966829	0.3647265	-0.44338140
-0.49123153	0.2127571	-0.07389684
1.14620239	0.6686652	-1.36709254
0.79024323	0.5166959	0.66507207
-1.34554581	-1.7628446	1.21929870

distance	rank
2.973764	2
2.511018	1
4.337104	4
3.172549	3

air.temp	humidity	uwind	vwind
NA	NA	-3	1.9

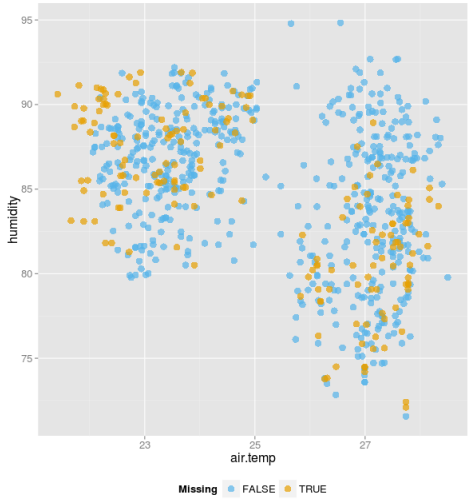
air.temp	humidity	uwind	vwind
23.54	92.2	-2.5	0.9
23.43	89.9	-2.7	1.1
23.89	88.6	-2.1	0.4
23.79	91.6	-2.3	1.5
NA	NA	-3.0	1.9

uwind	vwind
0.05725977	-0.4539797
-0.51533863	-0.1047645
1.20245657	-1.3270176
0.62985817	0.5936658
-1.37423589	1.2920960

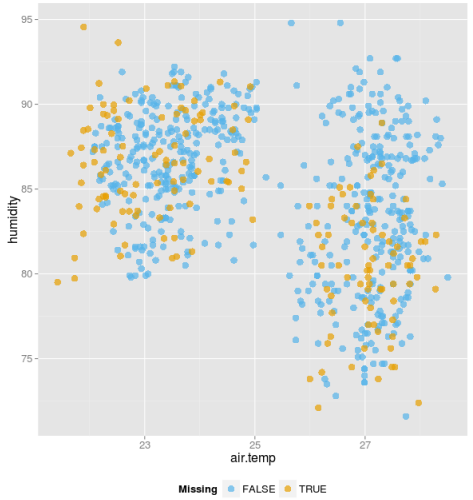
distance	rank
2.257866	3
1.639794	1
3.674112	4
2.122310	2

Multiple imputation

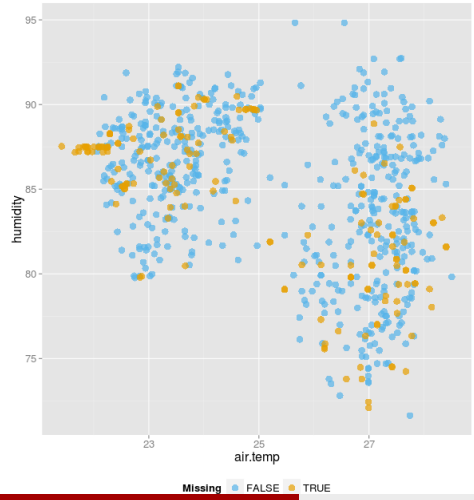
(a) `Hmisc`: predictive mean matching



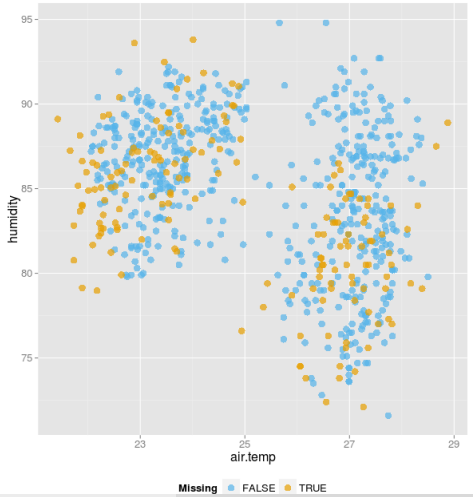
(b) `norm`: multivariate normal model



(c) `mice`: chained equations



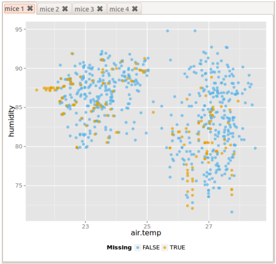
(d) `mi`: iterative regression

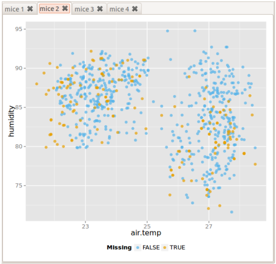


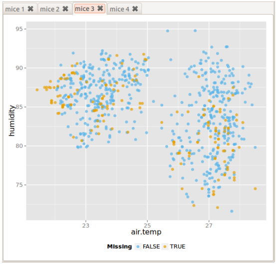
Multiple imputation

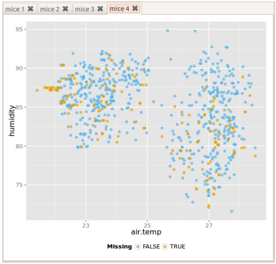
Algorithm Steps	Hmisc	mice	mi
1. fill in the missing	at random		
2. specify the model	pmm/regression/normpmm	selectable model or user-specific model	
(default model)	predictive mean matching		Baysian generalized linear models
3. decide the data	a bootstrap sample	the entire dataset with the current imputed values	
4. iterate imputation	in every cycle, variables with missings are imputed sequentially		
5. stop when	achieving the max # of iterations		difference of within and between variance is small

Multiple imputation

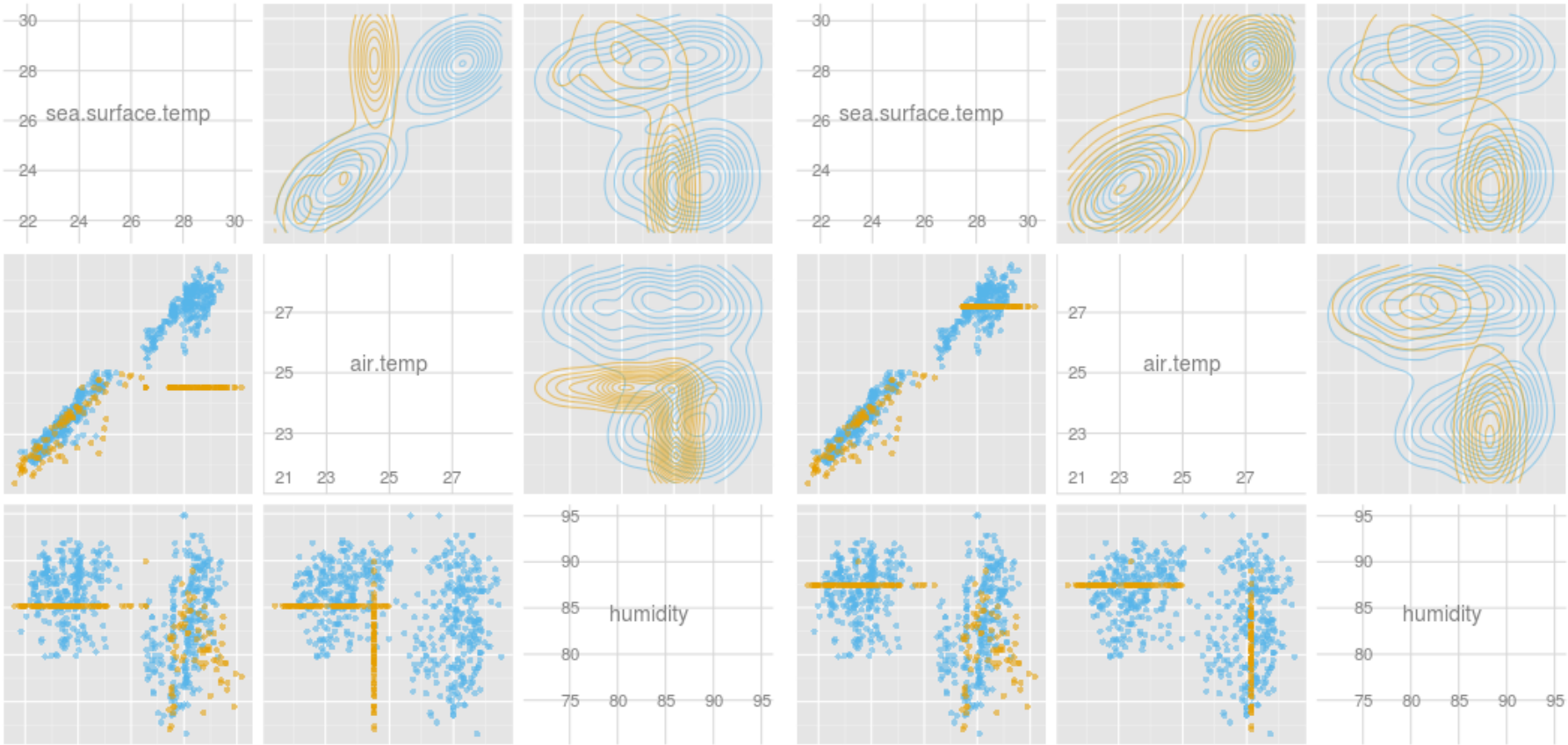






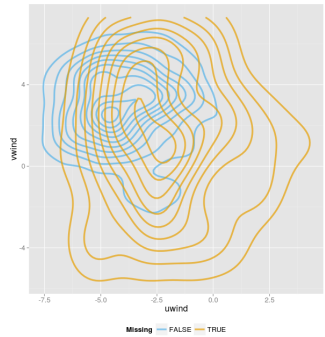


Condition on the categorical variables

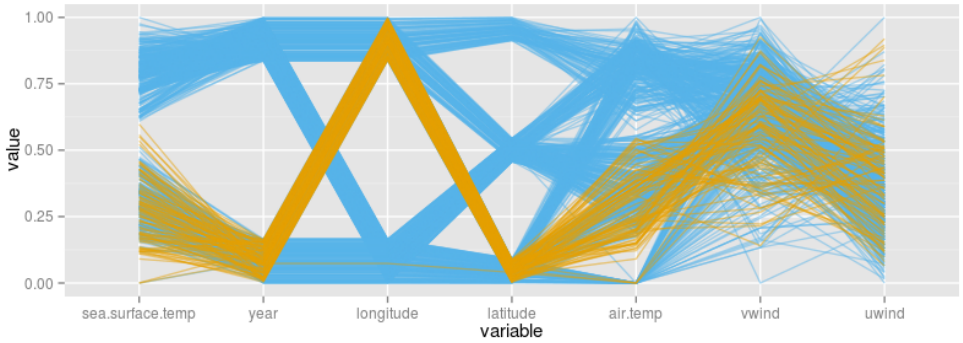


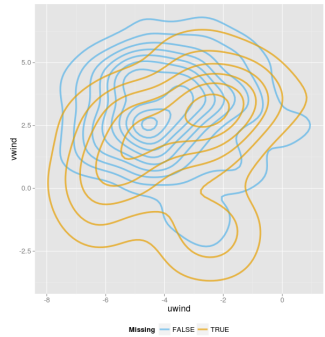
Check the assumptions

Missing on air.temp



Missing on humidity





Elements of Interactive Visualization for Temporal and Longitudinal Data

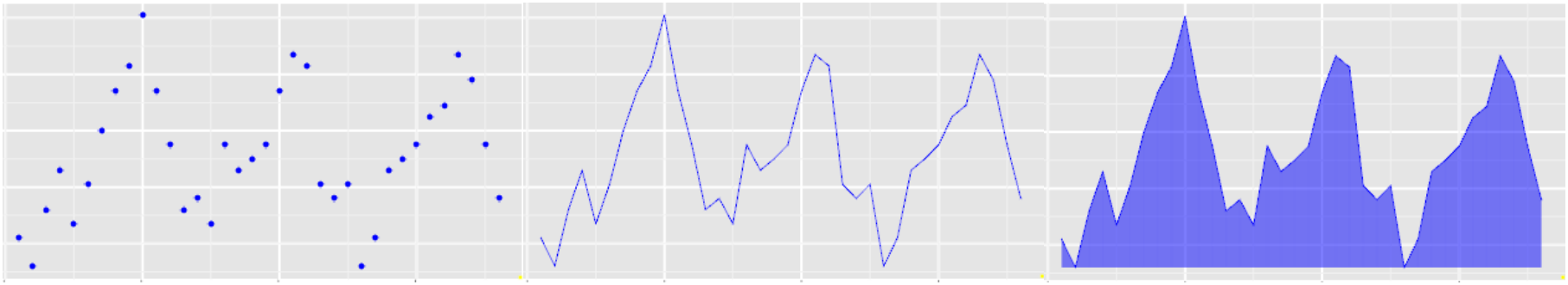
GUI vs. interactive graphics

- Two levels of interactions
  - human command: one-way, direct manipulation like GUI
  - human-computer cooperation: bi-direction with timely and consistent updates

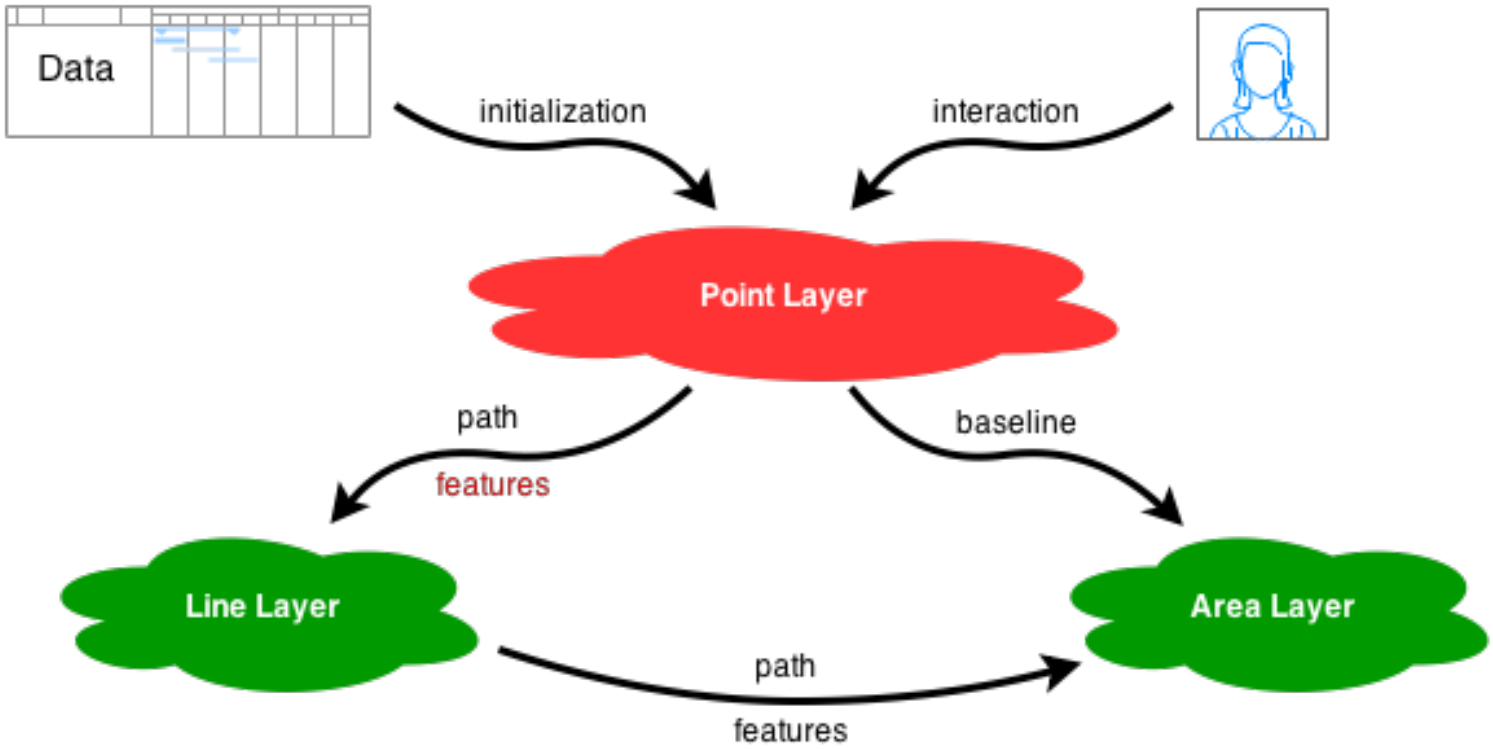
Elementary layers for the interactive graphics

common layers	required	brush, identify, keys
	optional	grid, x-axis, y-axis, x-label, y-label, title
special layers	scatterplot	point
	histogram	bar, cue
	map	polygon, googlemaps, path, point
	time series plot	point, line, area, stats

Layers for time visualization



Flow between the layers

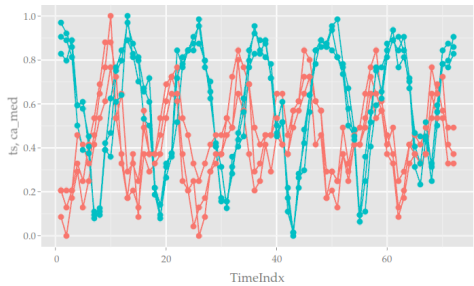


Interactivity

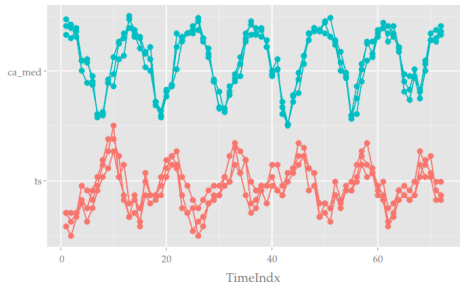
- Common interactions
  - brushing, selection, linking, zooming, panning, querying, etc.
- Special interactions for time series and longitudinal data
  - facetting
  - slicing and wrapping
  - mirroring
  - shifting

Facetting

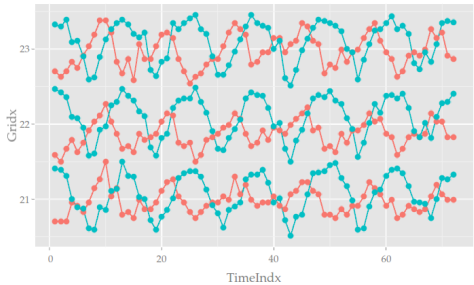
(a) not faceting



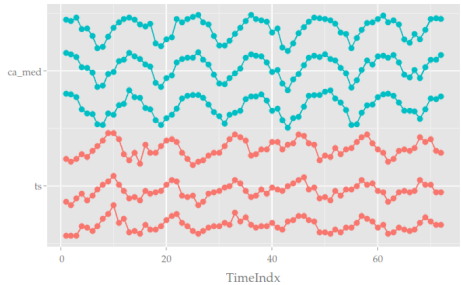
(b) faceting by variable



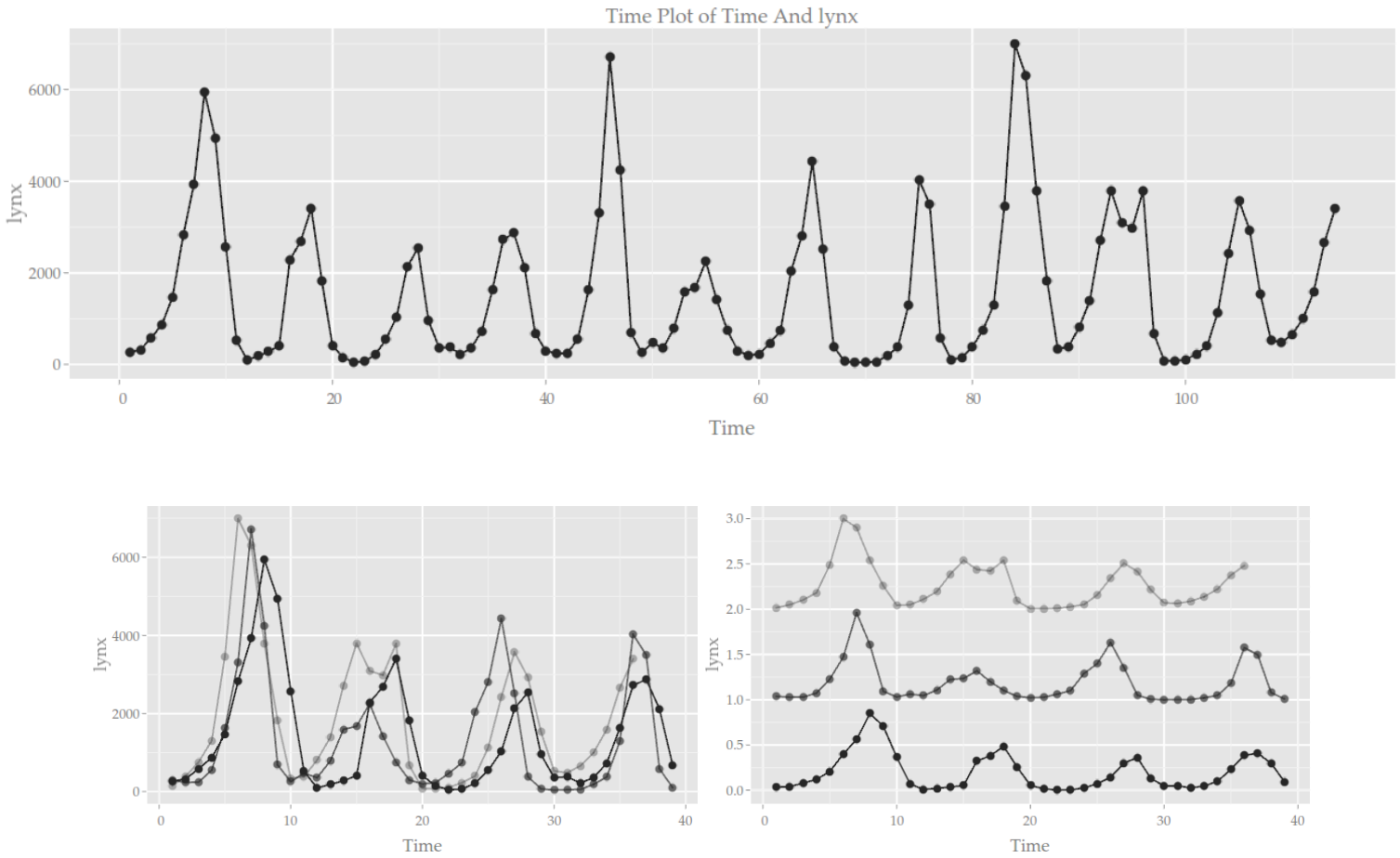
(c) faceting by individual



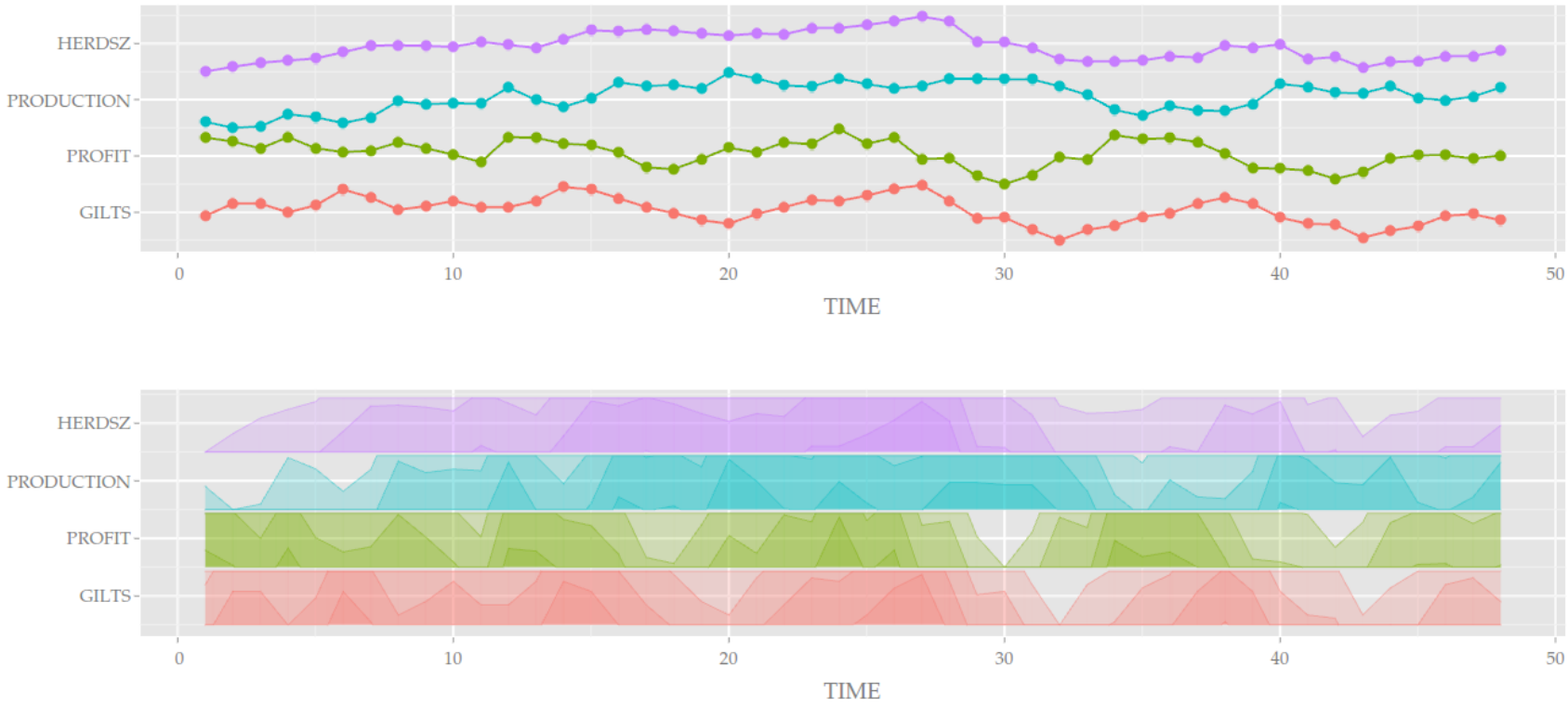
(d) by variable and individual



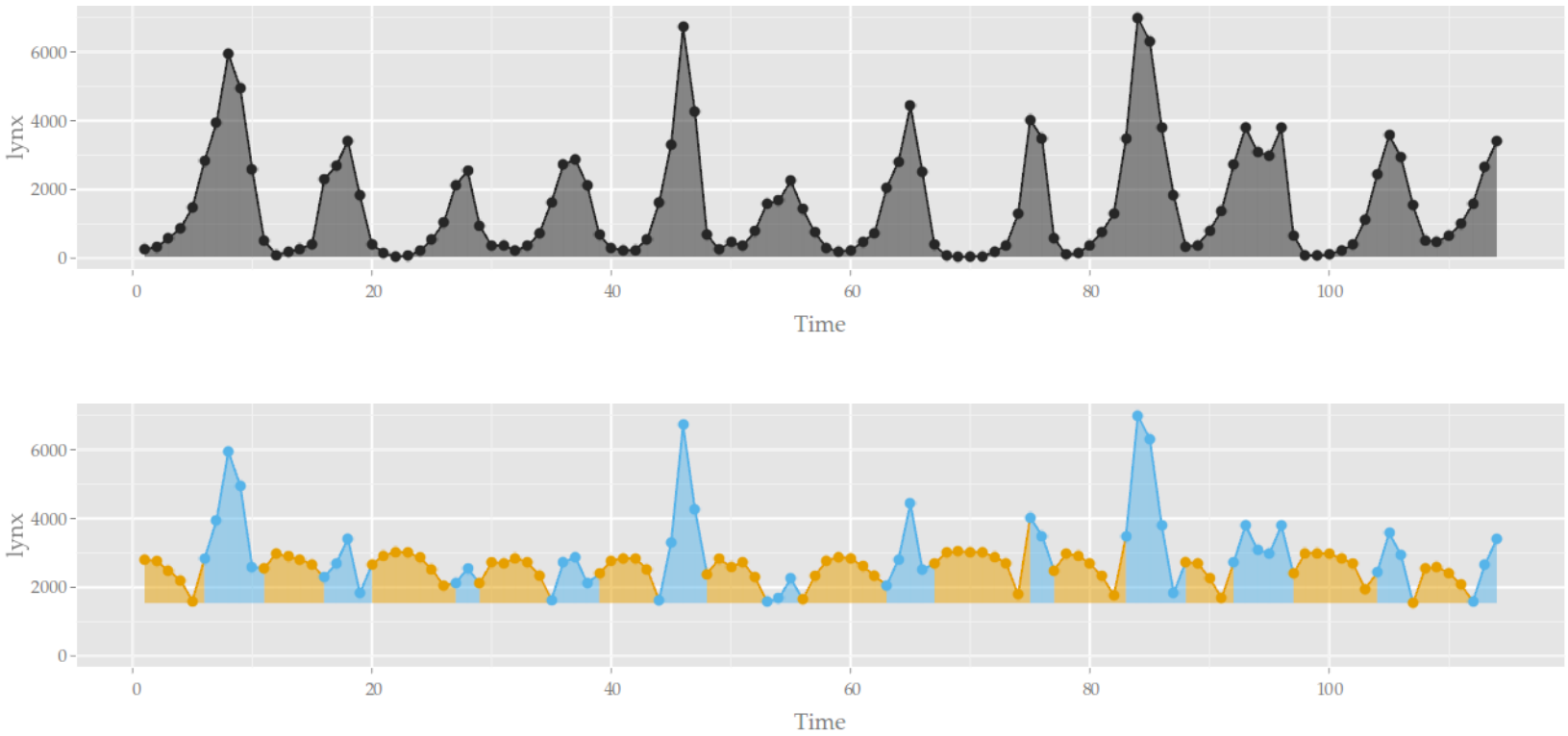
Slicing and wrapping



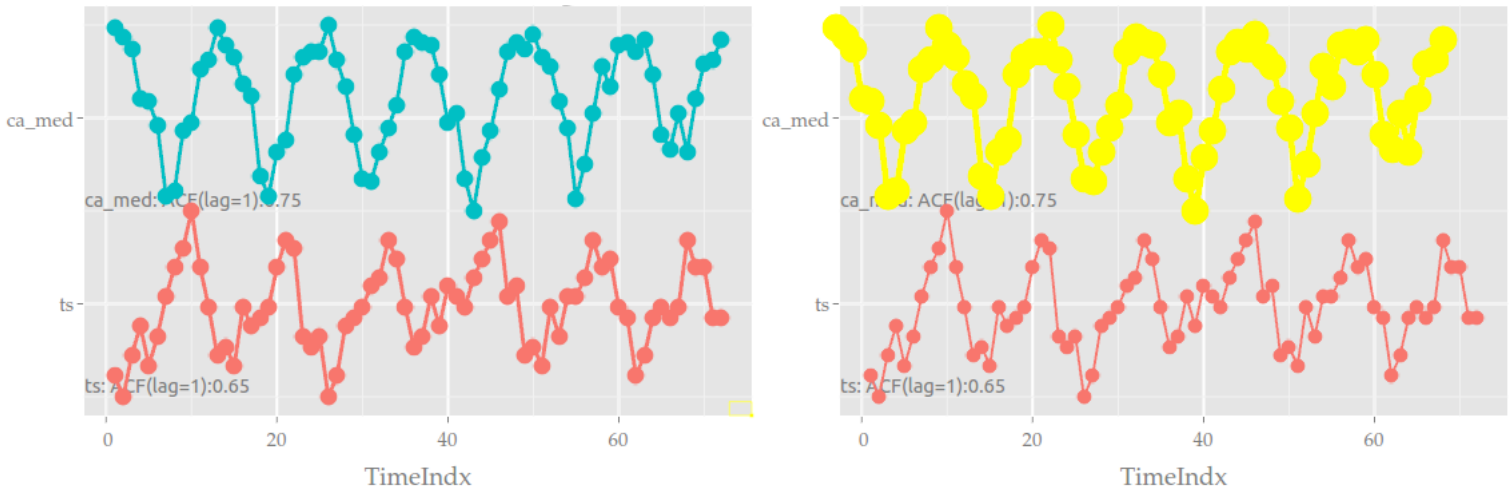
Slicing and wrapping



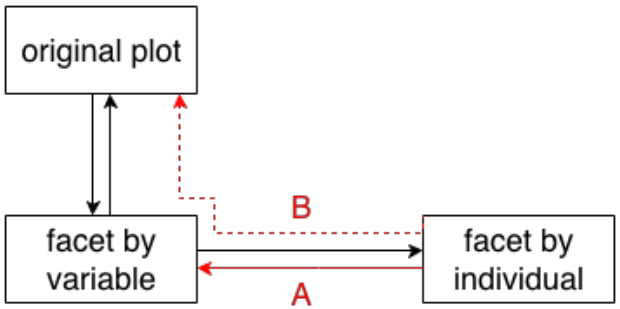
Mirroring


6000
4000
2000
0
lynx
0
20
40
60
80
100
Time
6000
4000
2000
0
lynx
0
20
40
60
80
100
Time

Shifting


ca_med
ca_med: ACF(lag=1):0.75
ts
ts: ACF(lag=1):0.65
0
20
40
60
TimeIndx
ca_med
ts
ts: ACF(lag=1):0.65
0
20
40
60
TimeIndx

Additivity of interactions



- Interactions are additive;
- Initial positions are needed for each interaction group.

Pipeline for parallel interactions

- Two designs
  1. current data = previous data + current interaction

$$data_s = f_s(data_{s-1},\ parameters_s)$$

$$parameters_s = g_s(parameters_{s-1},\ interaction_s)$$

  2. current data = original data + $\sum$interactions

$$data_s = f'_s(data_0,\ parameters_s)$$

$$parameters_s = g'_s(parameters_{s-1}, \cdots, parameters_0)$$

$$g'_s = interaction_1 + interaction_2 + \cdots + interaction_s$$

Linking

- Regular linking: on the same data, between graphs
- Additional linking: on different data, within and between graphs

Summary

- This work delineates the specific elements for temporal and longitudinal data visualization:
  - three basic layers
  - four special interactions
  - two designs of data pipeline for additive interactions
  - additional linking

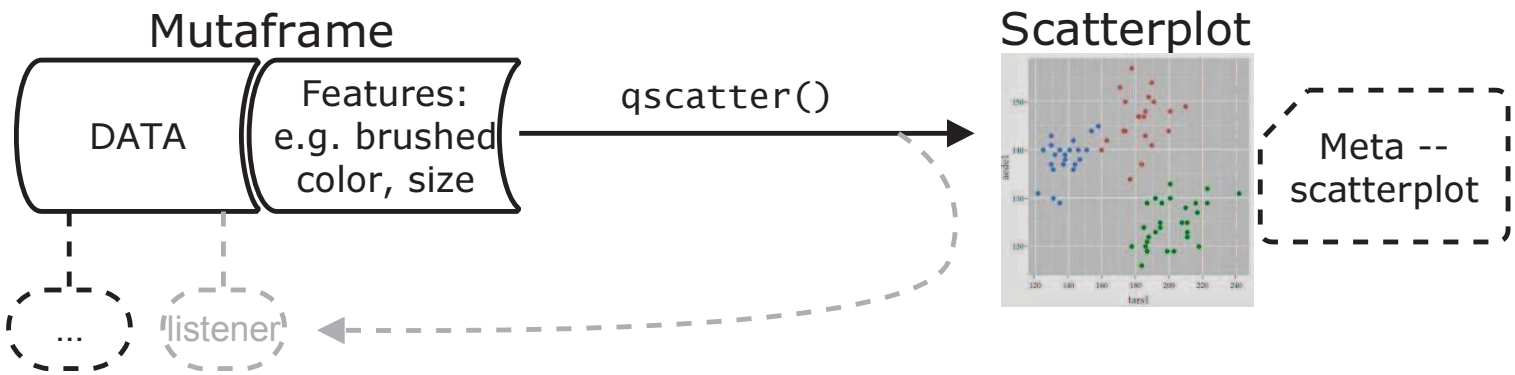
Cranvastime: An Interactive Tool for Temporal and Longitudinal Data Plots

Existing work

- Diamond Fast (Unwin and Wills, 1988)
- XQz(McDougall and Cook, 1994)
- XGobi and GGobi(Cook and Buja, 1997)
- Fortune(Kotter and Theus)
- d3.js(Bostock, 2012)

Cranvastime

- A part of the R package cranvas:



- A interactive visualization tool for uni-/multi-variate time series and longitudinal data
- Implementation of the grammar in Chapter 3

Example: univariate time series

- Lynx data: annual numbers of lynx trappings for 1821–1934 in Canada
- Sunspots data: monthly numbers of sunspots for 1749-1983

Example: multivariate time series

- NASA data: geographic and atmospheric measures on a very coarse 24 by 24 grid covering Central America
  - Obtained from the NASA Langley Research Center Atmospheric Sciences Data Center
  - Variables: elevation, temperature (surface and air), ozone, air pressure, and cloud cover (low, mid, and high)
  - Monthly averages from Jan 1995 to Dec 2000
- Pig data: quarterly production and profits for raising UK pigs during 1967-1978

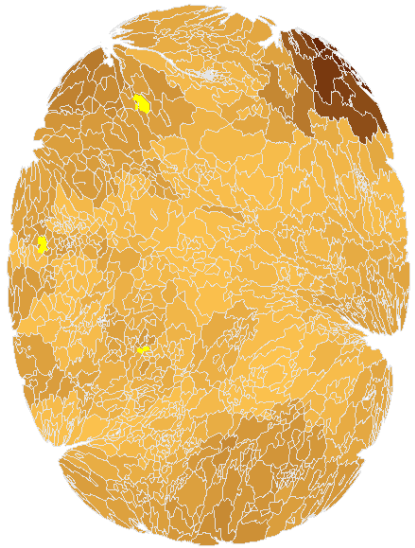
Example: longitudinal data

- Remifentanil data: pharmacokinetics of the drug, in R package `nlme`
- Google flu trends: daily flu-related searches

Cartogram Algorithms for Exploring Areal Data

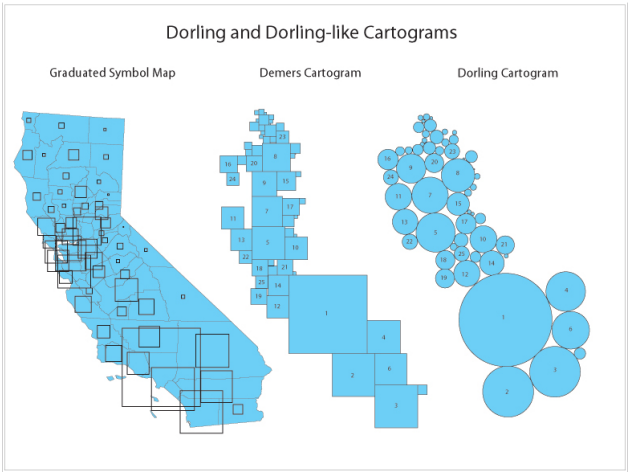
Motivation

- Experience at Amazon: need an area map on a zipcode level
- Cartogram is better than choropleth map
- Only one package in R, which gives



Existing Work

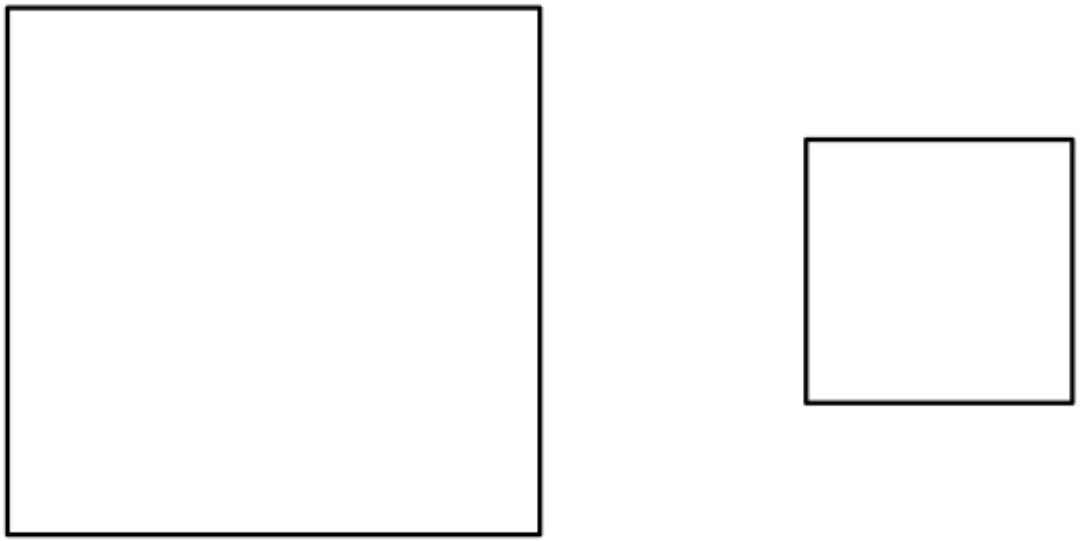
- Non-contiguous cartogram
  - shape-reserved method
  - Dorling cartogram



- Contiguous cartogram
  - rubber map
  - radial expansion
  - rubber sheet distortion
  - pseudo-cartogram
  - interactive polygon zipping
  - cellular automata machine
  - line integral
  - diffusion-based

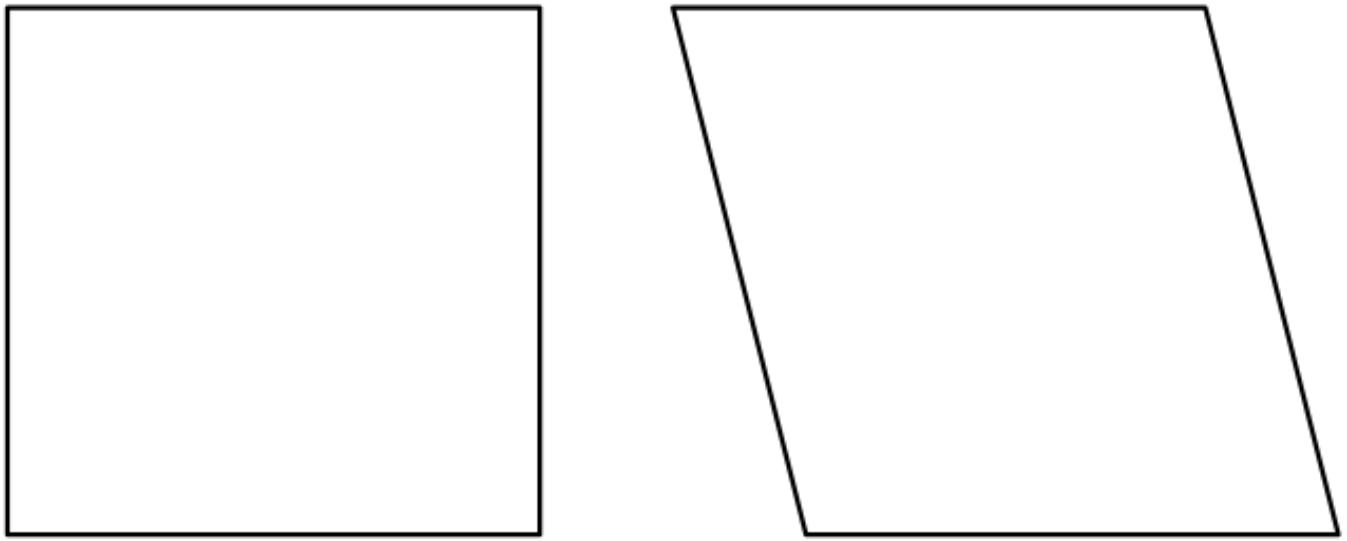
Assessment of a cartogram

- Measurement of three indexes
  - area: in proportion to the variable
  - position: fixed neighbors and relative directions
  - shape: similar to the original shape



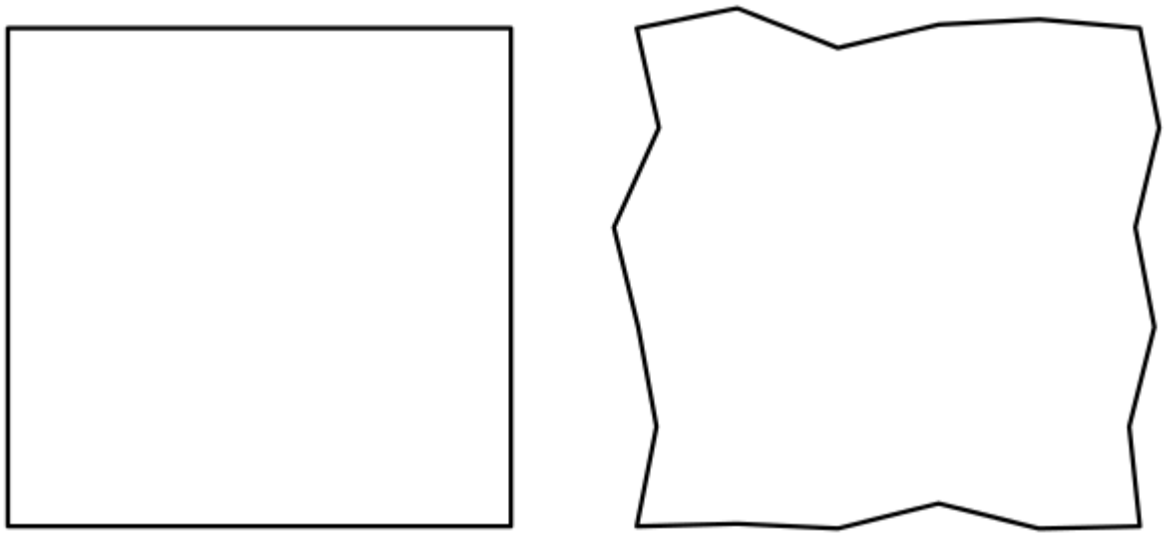
Assessment of a cartogram

- Measurement of three indexes
  - area: in proportion to the variable
  - position: fixed neighbors and relative directions
  - shape: similar to the original shape



Assessment of a cartogram

- Measurement of three indexes
  - area: in proportion to the variable
  - position: fixed neighbors and relative directions
  - shape: similar to the original shape

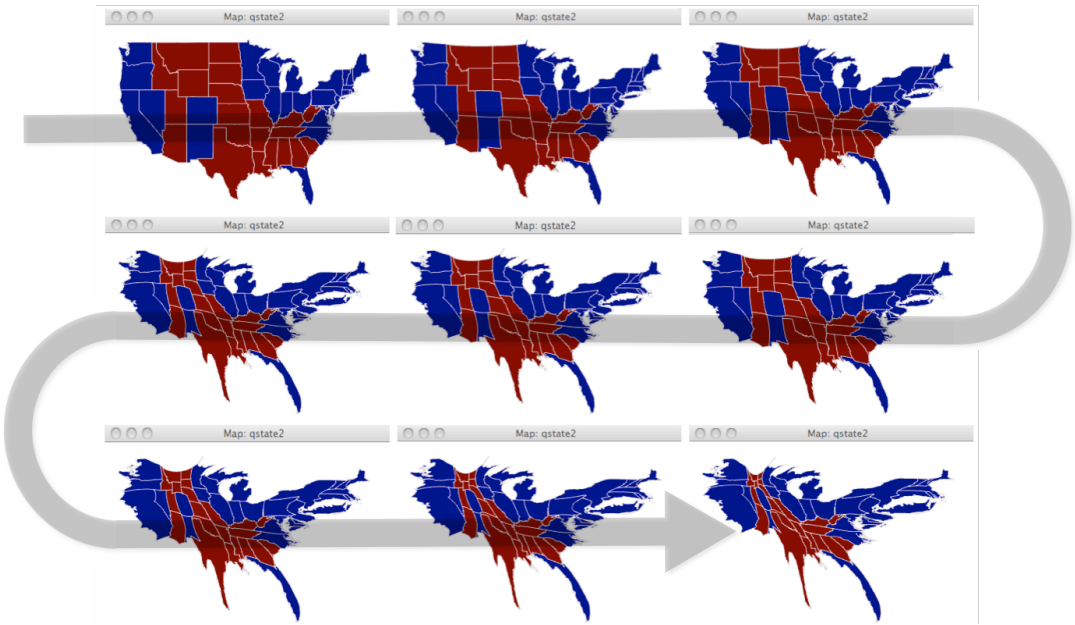


Method proposed

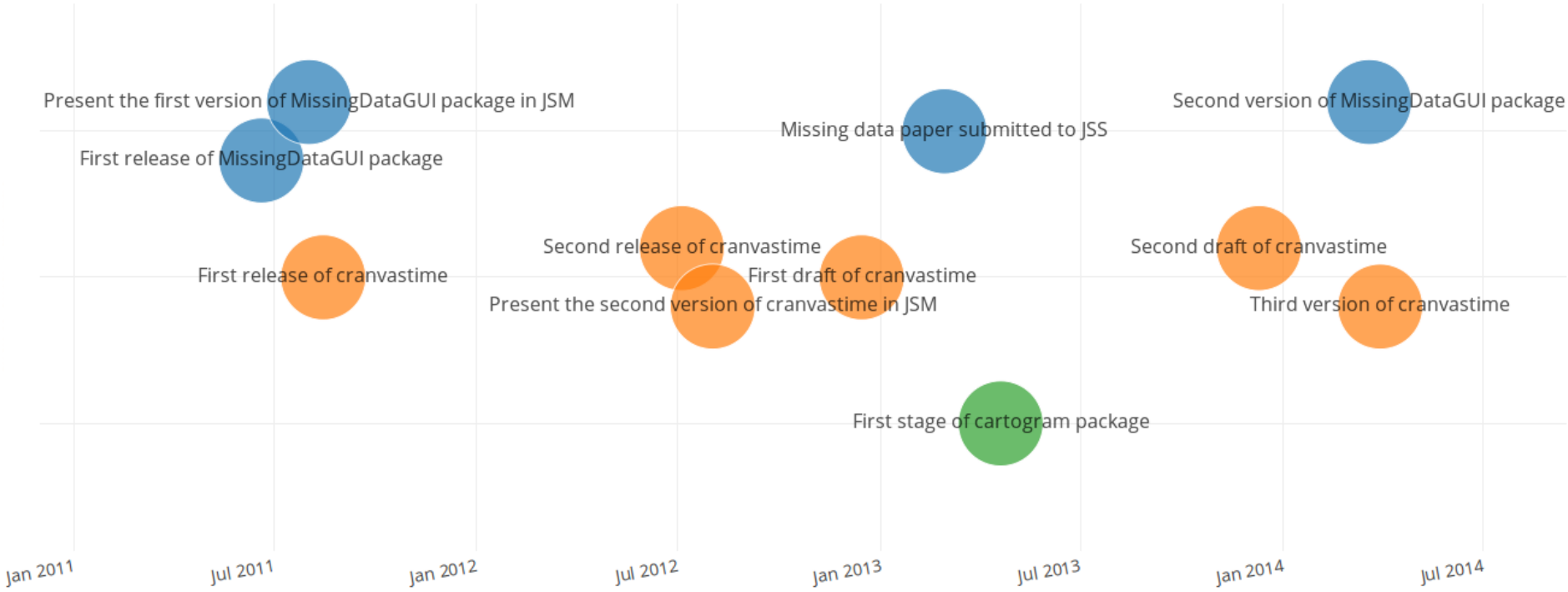
1. Start from non-contiguous shape-reserved cartogram
2. Centralize the polygons with the distance ratios remaining the same for each polygon
3. Predict the points in blank areas restricted to the position and distortion to the shape of new polygons

Interactiviy

- Like a choropleth map: brushing, selection, linking, zooming
- Shrinking



Completed Work



Scheduled deliverables

Product	Description	Date
R package	First release of the `cartogram` package	May 2014
Paper	Pipeline for the interactive visualization on temporal and longitudinal data	July 2014
Talk	Present the areal visualization in JSM	August 2014
Paper	Interactive visualization on areal data	October 2014
Defence	Thesis defence	November 2014

Other Work

Product	Description	Date
R package	First release of the `MergeGUI` package, which can merge multiple data sets by checking the consistency of variables.	June 2011
Technical document	Manual for the `cranvas` package	August 2013
R package	Second release of the `MergeGUI` package	January 2014
Technical report	Reproducible reports on introductory statistics eduation	February 2014
Paper	Reactive Programming for Interactive Graphics. Joint work with Yihui Xie and Heike Hofmann. To Appear in Statistical Science.	March 2014

References

Bostock, M. (2012). Data-driven documents (d3. js), a visualization framework for internet browsers running javascript.

Brix, P. (2012). *miP: Multiple Imputation Plots*. R package version 1.1.

Buja, A., Asimov, D., Hurley, C., and McDonald, J. A. (1988). Elements of a viewing pipeline for data analysis. *Dynamic graphics for statistics*, pages 277–308.

Cleveland, W. S. and McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society. Series A (General)*, pages 192–229.

Cook, D. and Buja, A. (1997). Manual controls for high-dimensional data projections. *Journal of Computational and Graphical Statistics*, pages 464–480.

Fisherkeller, M. A., Friedman, J. H., and Tukey, J. W. (1988). Prim-9, an interactive multidimensional data display and analysis system. *Dynamic Graphics for Statistics*, pages 91–109.