Visual Data Mining in Education, Social Sciences, and Environmental Sciences

Jürgen Symanzik

Utah State University, Logan, UT, USA

e-mail: symanzik@math.usu.edu

http://www.math.usu.edu/~symanzik

HU Berlin, IRTG Short Course (II)

July 20, 2016

Contents

- Terms, Citations, and Definitions
- Case Study 1: An iPad Study in Education
- Case Study 2: The “Soul of the Community” Social Sciences Project
- Case Study 3: Forecasting of “Snow Water Equivalent” Measurements
- Case Study 4: Exploratory Graphics for Functional Actigraphy Data in Sleep Medicine
- Conclusion

Terms

- Interactive & Dynamic Statistical Graphics (DSG)
- Exploratory Data Analysis (EDA)
- Exploratory Spatial Data Analysis (ESDA)
- Visual Data Mining (VDM)
- Visual Analysis/Visual Analytics (VA)
- Data Mining (DM)

Citations

- John W. Tukey (1977):

  EDA *"is detective work - numerical detective work - or counting detective work - or graphical detective work."*

- Edward J. Wegman (2000):

  "Data Mining is exploratory data analysis with little or no human interaction using computationally feasible techniques, i.e., the attempt to find interesting structure unknown a priori."

Interview with Andreas Buja*

- " … when I think back about what really may have had the most impact in what I did in the various labs that I worked, it's graphics! You know whenever I made a striking picture, people actually went "aahh," "wow," "that's great!", "Why don't we do more of this?" Pictures really, really speak. …"

- *Computational Statistics (2008) 23:177–184

Visual Data Mining (1)

- Working Definition for VDM:
  - Find structure (cluster, unusual observations) in large and not necessarily homogeneous data sets based on human perception using graphical methods and user interaction
  - Goal or expected outcome of exploration usually unknown in advance

Visual Data Mining (2)

- First uses of the term VDM:
  - – Cox, Eick, Wills, Brachman (1997): Visual Data Mining: Recognizing Telephone Calling Fraud, *Data Mining and Knowledge Discovery*, 1:225-231.
  - – Inselberg (1998): Visual Data Mining with Parallel Coordinates, *Computational Statistics*, 13(1):47-63.

Visual Data Mining Concepts

- Use existing visualization techniques, such as
  - Scatterplots and Scatterplot Matrices
  - Parallel Coordinate Plots
  - Heatmaps
  - Mosaic Plots
  - Brushing and Linked Brushing/Linked Views
  - Rotations and Projections
  - Grand Tour
  - “Small Multiples”, …
- Develop customized visualization techniques

Main References

- Symanzik, J. (2012): Interactive and Dynamic Graphics [Revised], In: Gentle, J. E., Härdle, W. K., Mori, Y. (Eds.), Handbook of Computational Statistics --- Concepts and Methods, Vol. 1 (Second Revised and Updated Edition), Springer, Berlin/Heidelberg, 335-373.

- Symanzik, J. (2011): Interactive and Dynamic Statistical Graphics, In: Lovric, M. (Ed.), International Encyclopedia of Statistical Science, Springer, Berlin/Heidelberg, 674-679.

Case Study 1: An iPad Study in Education

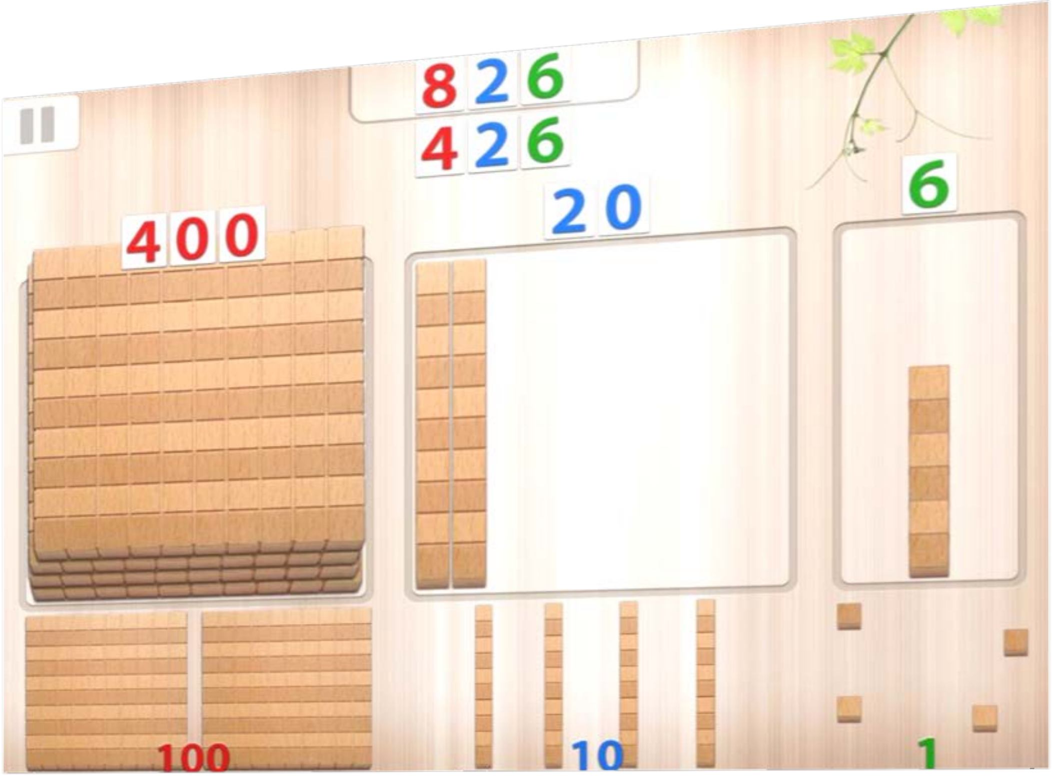
Published as:

Moyer-Packenham, P. S., Shumway, J. F., Bullock, E., Tucker, S. I., Anderson-Pence, K. L., Westenskow, A., Boyer-Thurgood, J., Maahs-Fladung, C., Symanzik, J., Mahamane, S., MacDonald, B., Jordan, K. (2015): Young Children's Learning Performance and Efficiency when Using Virtual Manipulative Mathematics iPad Apps, *Journal of Computers in Mathematics and Science Teaching*, 34 (1): 41-69.

Moyer-Packenham, P. S., Tucker, S. I., Westenskow, A., Symanzik, J. (2015): Examining Patterns in Second Graders' Use of Virtual Manipulative Mathematics Apps through Heatmap Analysis, *International Journal of Educational Studies in Mathematics*, 2 (2): 1-16 .

Purpose of the Study

- The purpose of the project was to build theory and knowledge about the nature of young children's ways of thinking and interacting with virtual manipulative mathematics apps on the iPad.



Research Questions

Learning Performance and Efficiency:

- What are the immediate effects in learning performance and efficiency (pre vs. post) for children using virtual manipulatives for the iPad?

Learning Strategies:

- How do children interact with the virtual manipulatives on the iPad using the touch-screen capabilities?

Participants

- 100 children ages 3 to 8
  - 35 preschool, ages 3-4
  - 33 Kindergarten, ages 5-6
  - **32 Grade 2, ages 7-8**

- Demographic information were collected on age, gender, race, prior iPad use, etc.

Procedures: Clinical Interviews

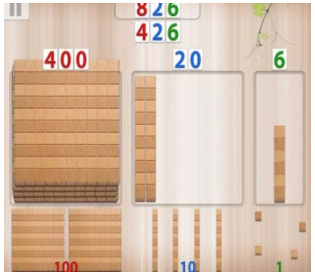
Sequence of the Interviews

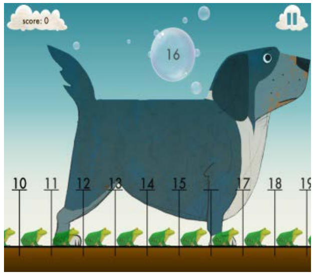
Interview	Grade Pre	Grade K	Grade 2
App #1 (pre)	Pink Tower: free moving	10-Frame	100s Chart
App #2 (learning)	Pink Tower: tapping	Hungry Guppy	Frog Number Line
App #3 (learning)	Red Rods	Fingu	Counting Beads
App #1 (post)	Pink Tower: free moving	10-Frame	100s Chart
App #4 (pre)	Base-10 Blocks	Base-10 Blocks	Base-10 Blocks
App #5 (learning)	Base-10 Blocks: 1-5	Base-10 Blocks: 11-20	Zoom Number Line
App #6 (learning)	Base-10 Blocks: numerals	Base-10 Blocks: numerals	Place Value Cards
App #4 (post)	Base-10 Blocks	Base-10 Blocks	Base-10 Blocks

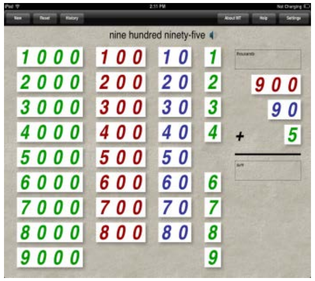
Grade 2 Interview Apps

Quantities

- Pre/Post
  - Base-10 Blocks
- Activity A
  - Math Motion Zoom (Levels 2-4)
- Activity B
  - Place Value Cards (3-digit problems without zeros)





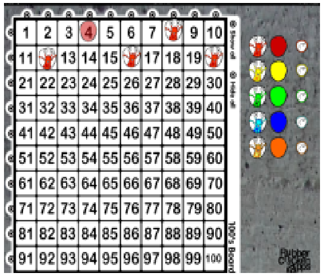

nine hundred ninety-five

1000	100	10	1
2000	200	20	2
3000	300	30	3
4000	400	40	4
5000	500	50	
6000	600	60	6
7000	700	70	7
8000	800	80	8
9000			9

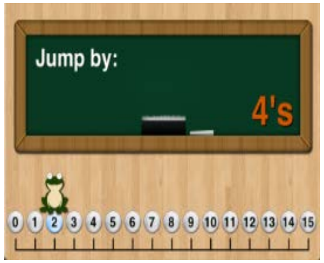
900
90
+ 5

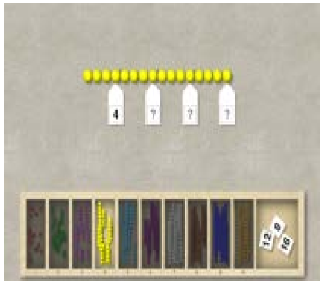
Skip Counting

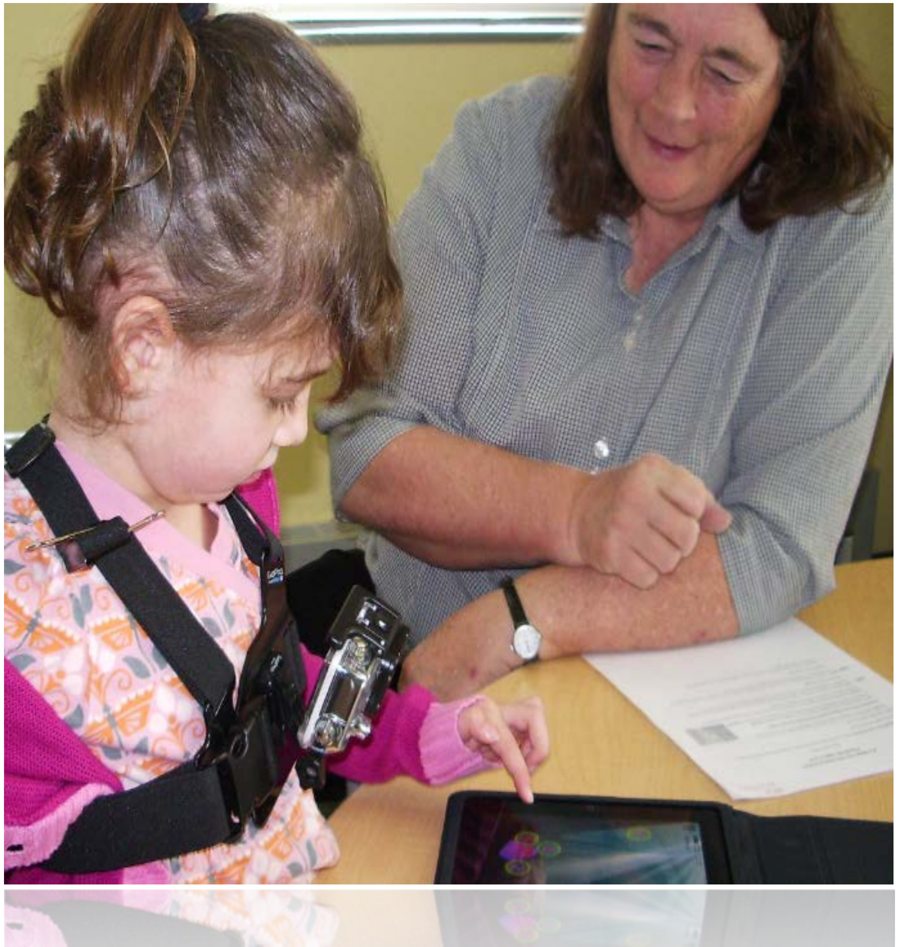
- Pre/Post
  - 100s Chart
- Activity A
  - Number Lines (Skip Counting Tool
- Activity B
  - Skip Counting Beads



1	2	3	4	5	6	7		9	10
11		13	14	15		17	18	19	
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100







Data Collection

Video Cameras:

- Wall-mounted
- Go-Pro

Data Analysis

- Base-10 Blocks: 6 performance variables (1: Model a number between 12 and 30; 2: Model a number between 54 and 62; 3: Model 181; 4: Model a number between 181 and 200; 5: Model 267; & 6: Model a number 20 less than 267)
- 100s Chart: 3 performance variables (skip counting by 4, 6 & 9)
- Performance data were scaled from 0 (very poor) to 1 (excellent)
- Analyzing learning performance from Pre- and Post-assessments

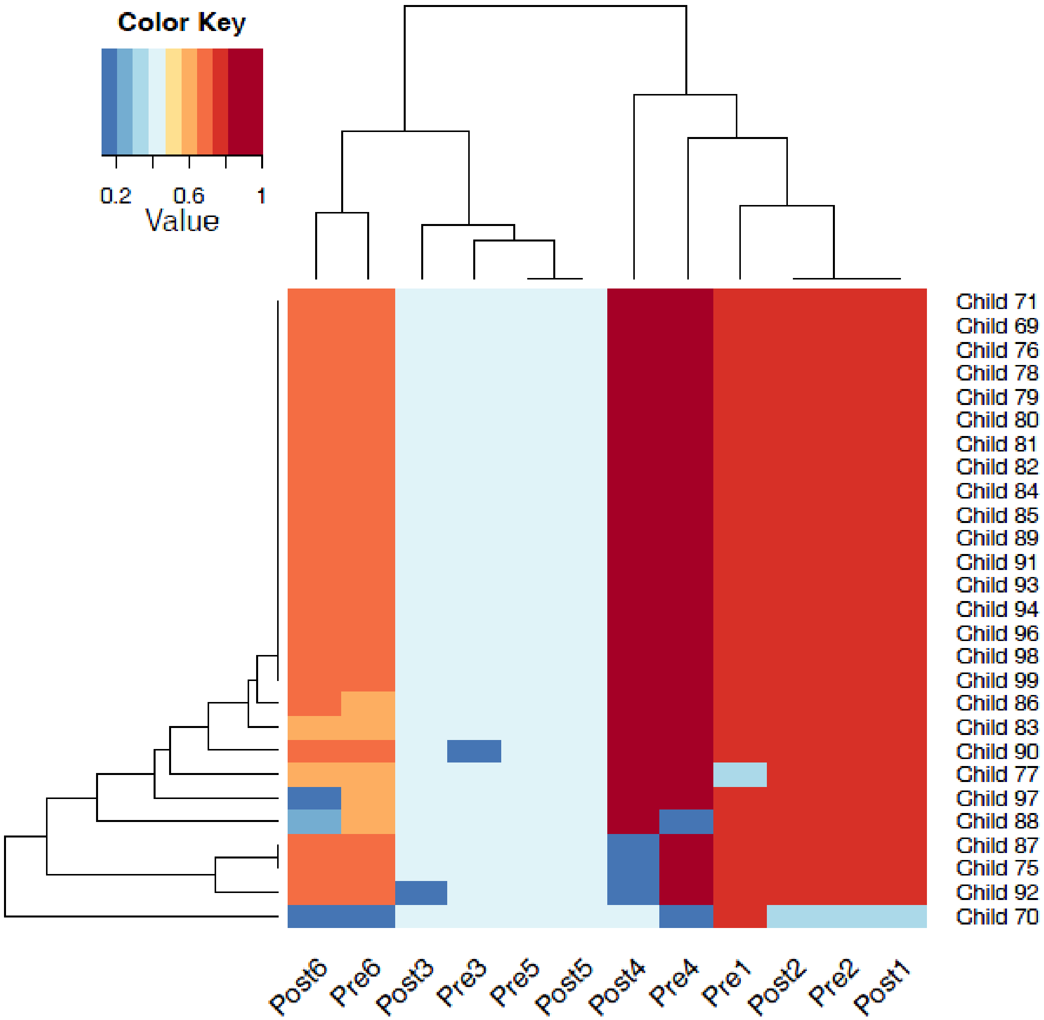
Grade 2: Base-10 Blocks

Base-10 Blocks (1)

- 17 of the 27 children had totally identical outcomes. Moreover, the outcomes for the two children # 75 and 87 were totally identical. Only eight children were more different. Child # 70 was most unusual with only a single score higher than 0.5.

- Scores from the six Pre and Post test pairs matched each other very closely.

- Variables Post1, Pre2, and Post2 had totally identical outcomes with a score of 0.875 for all but one child. Variable Pre1 matched them for all but two children.



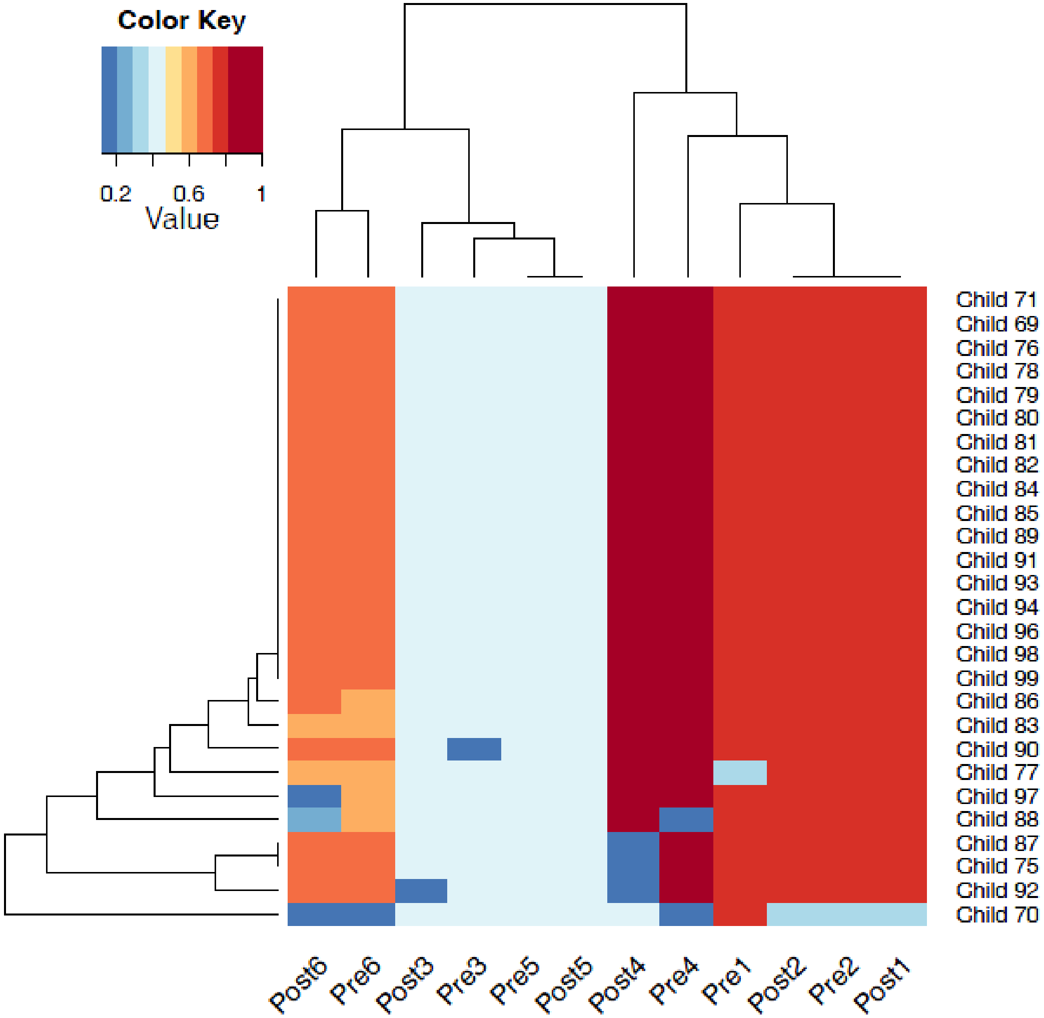
Base-10 Blocks (2)

- Variables Pre5 and Post5 had totally identical outcomes with a score of 0.5 for all children.

- Variables Pre3 and Post3 matched them for all but two children.

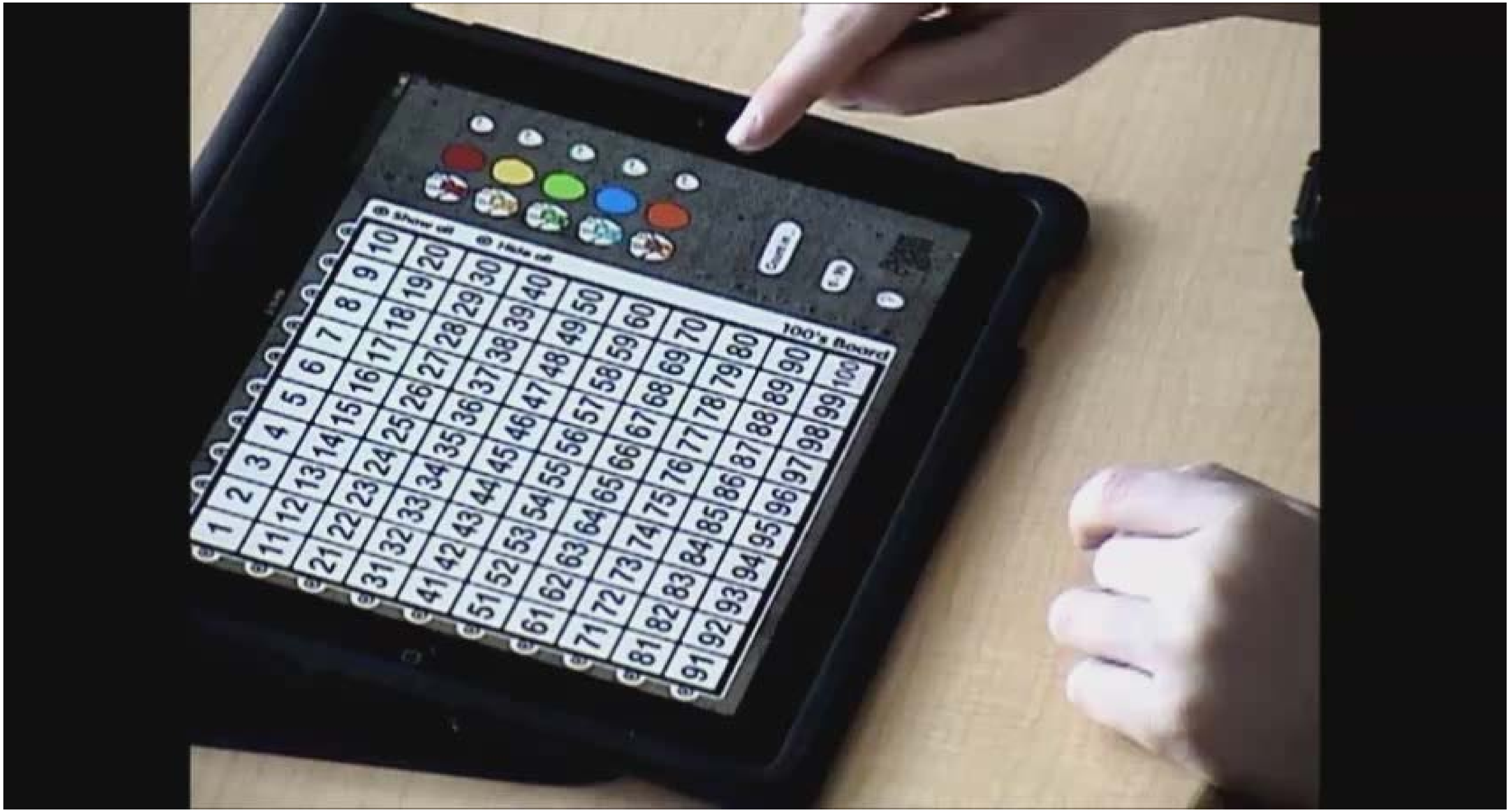
- Variables Pre6 and Post6 had almost identical outcomes with a score of 0.75 for most children. Only three children's outcomes were different in their pre/post tests.

- Variables Pre4 and Post4 had almost identical outcomes with a score of 1.0 for most children. Only five children's outcomes were different in their pre/post tests.



Grade 2: Skip Counting

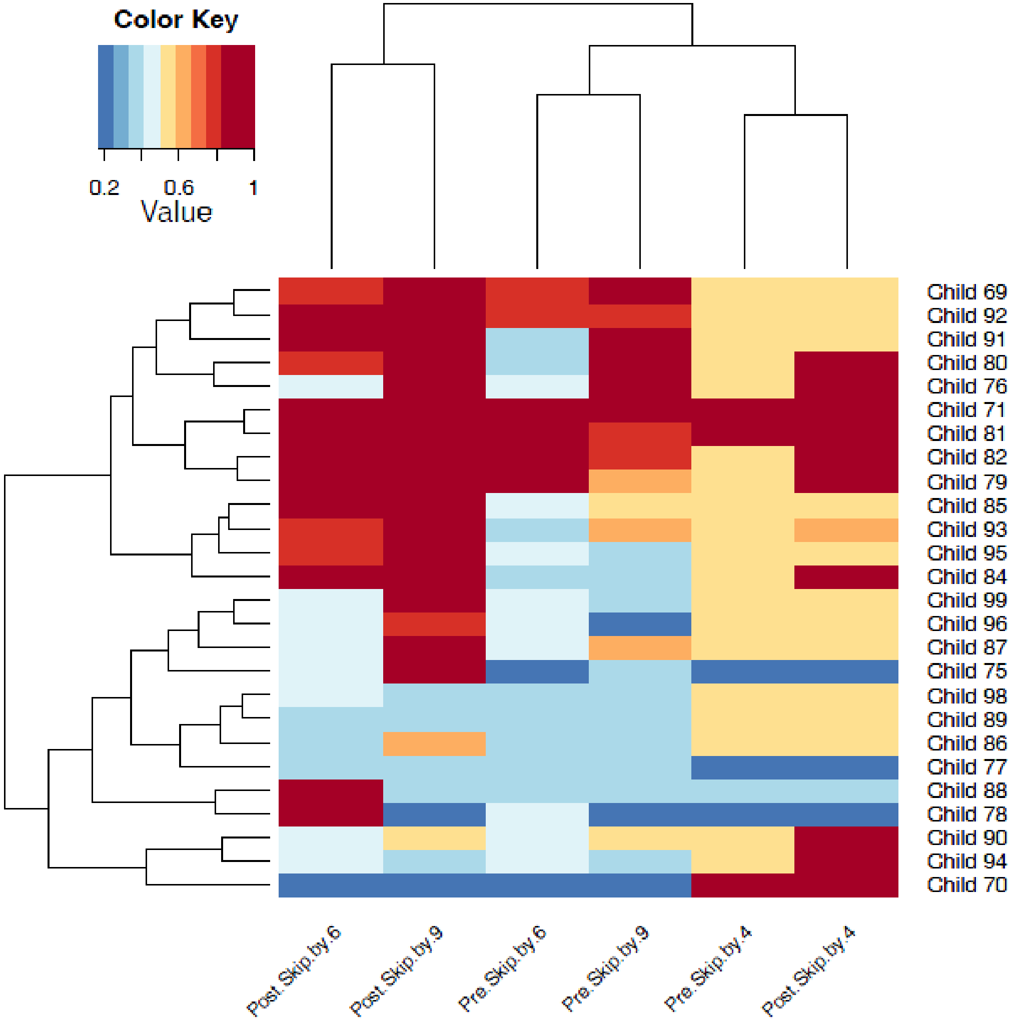
100s Chart



1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

100s Chart

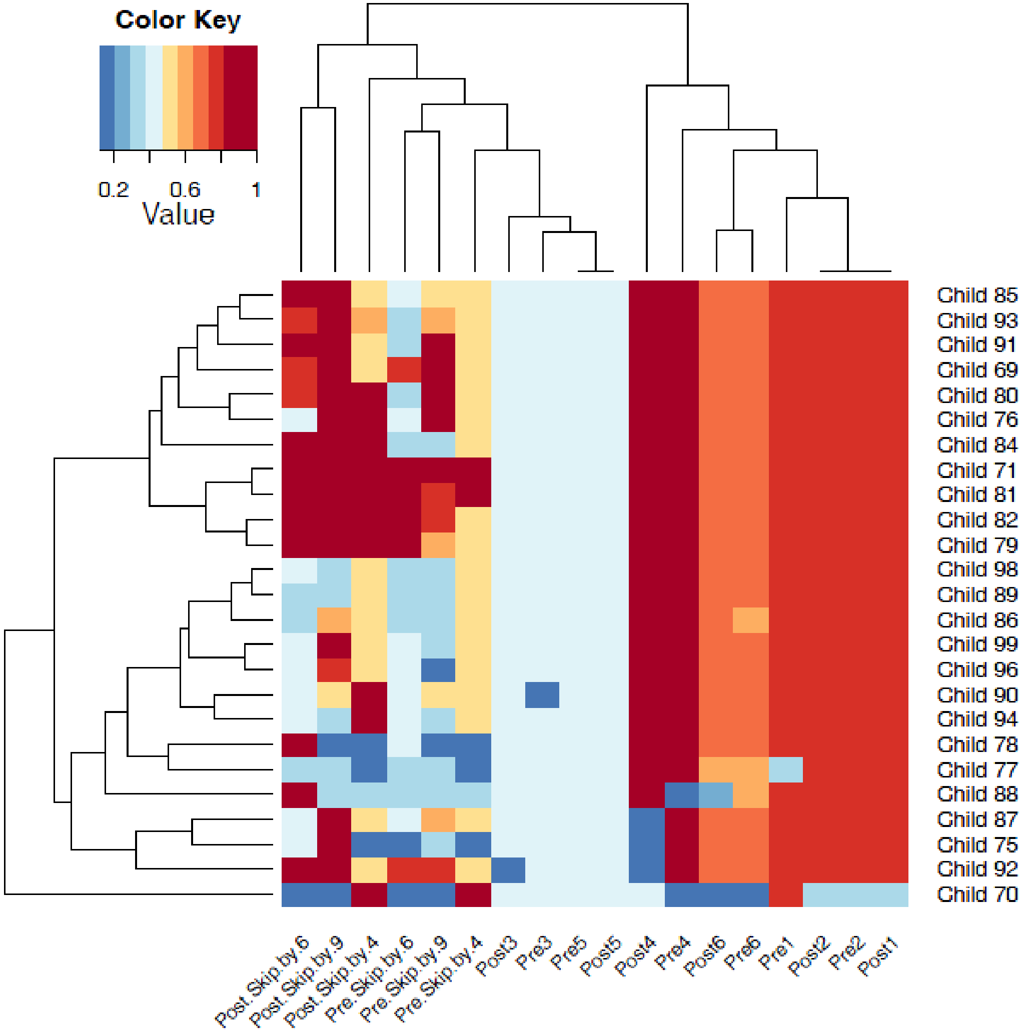
- In the dendrogram, there are three pairs of two variables: Pre.Skip.by.4 and Post.Skip.by.4, Pre.Skip.by.6 and Pre.Skip.by.9, and Post.Skip.by.6 and Post.Skip.by.9.
- There are no children who had exactly the same outcomes.
- There are two main clusters: In cluster 1 (13 children: # 69 to # 84 on the right), children had medium to high scores for most of the variables. In cluster 2 (13 children: # 99 to # 70 on the right), children had low to medium scores for most of the variables.
- Child # 71 is the only child with a perfect 1.0 for all six variables.



Summary (1)

- In the dendrogram, there are two main groups of variables: Pre/Post 1, 2, 4, and 6 (group 1) and Pre/Post 3 and 5, and Pre/Post skip.by 4, 6, and 9 (group 2). This means Pre/Post 3 and 5 are somewhat closer to the variables from 100 Chart than to the eight other variables from Base 10.

- There are no children who had exactly the same outcomes.

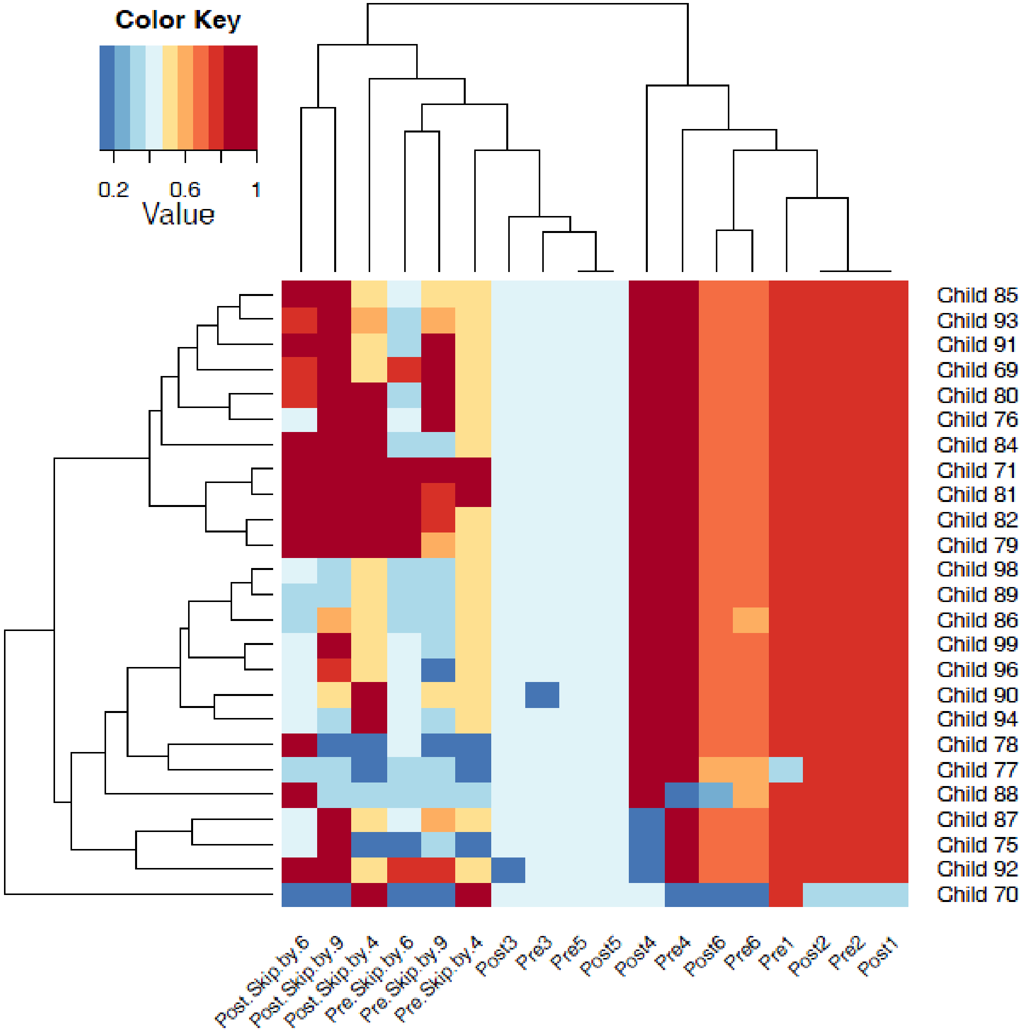


Summary (2)

- There are two main clusters: In cluster 1 (11 children: # 85 to # 79 on the right), children had medium to high scores for most of the variables. In cluster 2 (13 children: # 98 to # 92 on the right), children had low to medium scores for most of the variables.

- Child # 70 is the most unusual child that differs considerably from both main clusters of children.

- Second-grade children increased their performance on the skip counting app, but not on the base 10 app.



Case Study 2: The “Soul of the Community” Social Sciences Project

Published as:

Quach, A., Symanzik, J., Forsgren, N. (2016): Soul of the Community: An Attempt to Assess Attachment to a Community, *Computational Statistics*, Accepted.

Background (1)

- Soul of the Community Survey (SOTC) Project:
  - Conducted by the Knight Foundation
  - Time period: 2008 to 2010
  - 26 communities across the United States
  - More than 47,800 participants
  - Around 200 different questions each year
- Key variable: Attachment to one's community

Background (2)

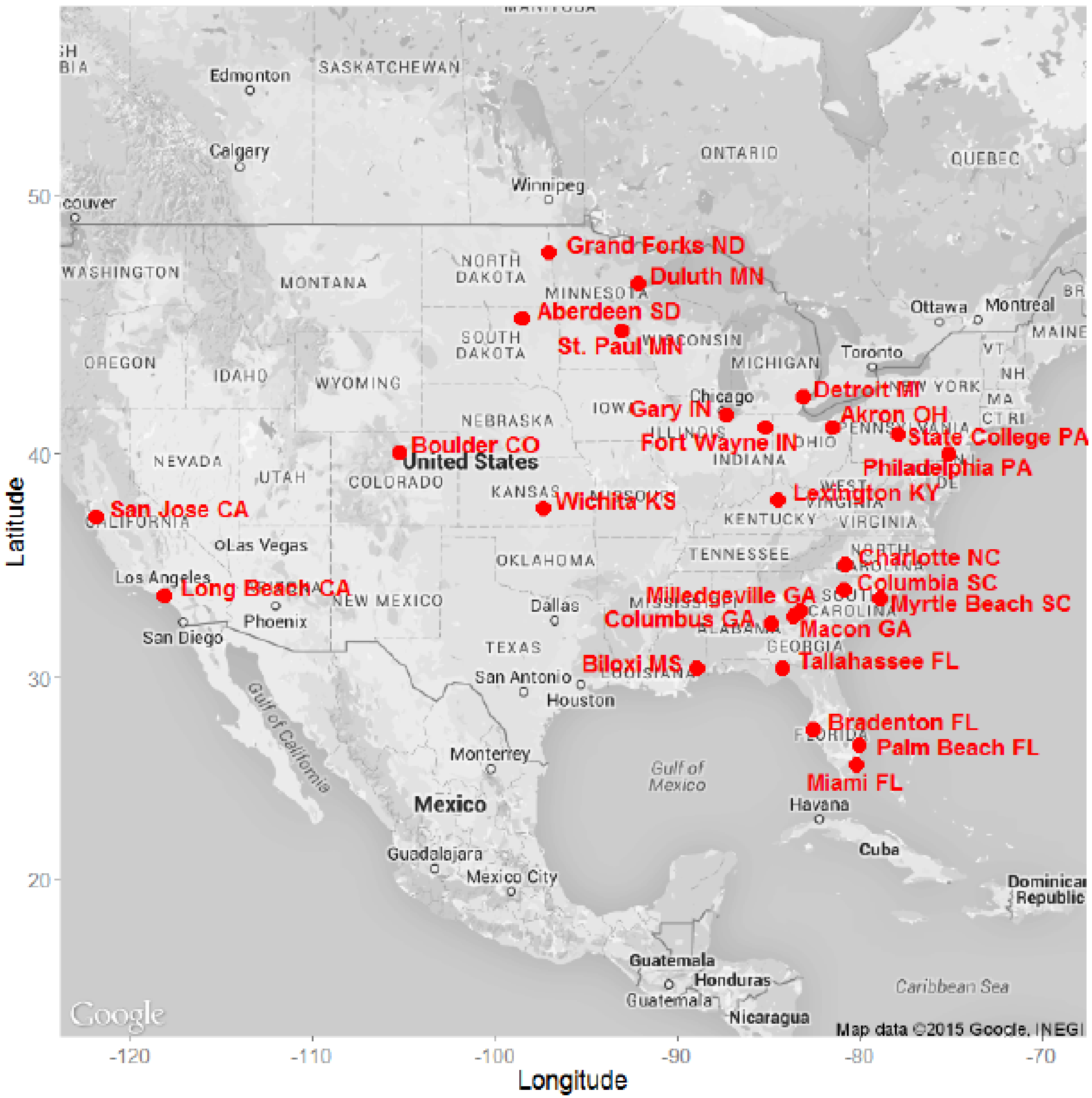


Fig. 1: Locations of the 26 communities involved in the SOTC project, overlaid on a Google map.

Questions

- Which factors foster attachment to one's community?
  - Which factors impact attachment to particular communities?
  - Which factors impact attachment to communities as a whole?
- Are there differences in attachment between communities as well as demographics?

Methods

- Random Forests (RF)
- Support Vector Machines (SVM)
- Multiple Linear Discriminant Analysis (LDA)
- Recursive Partitioning And Regression Trees (RPART)
- Archetypal Analysis

RF Results

- Variables:
  - q3c: The community has a good reputation to outsiders or visitors who do not live here: 1 - Strongly disagree … 5 - Strongly agree
  - q5: If you had the choice of where to live would you rather: 1 - stay in your neighborhood; 2 - move to another neighborhood; 3 - Move outside of your community; 4 - Move to another city and state
  - q6: How would you compare how the community is as a place to live today compared to five years ago: 1 - Much worse … 5 - Much better

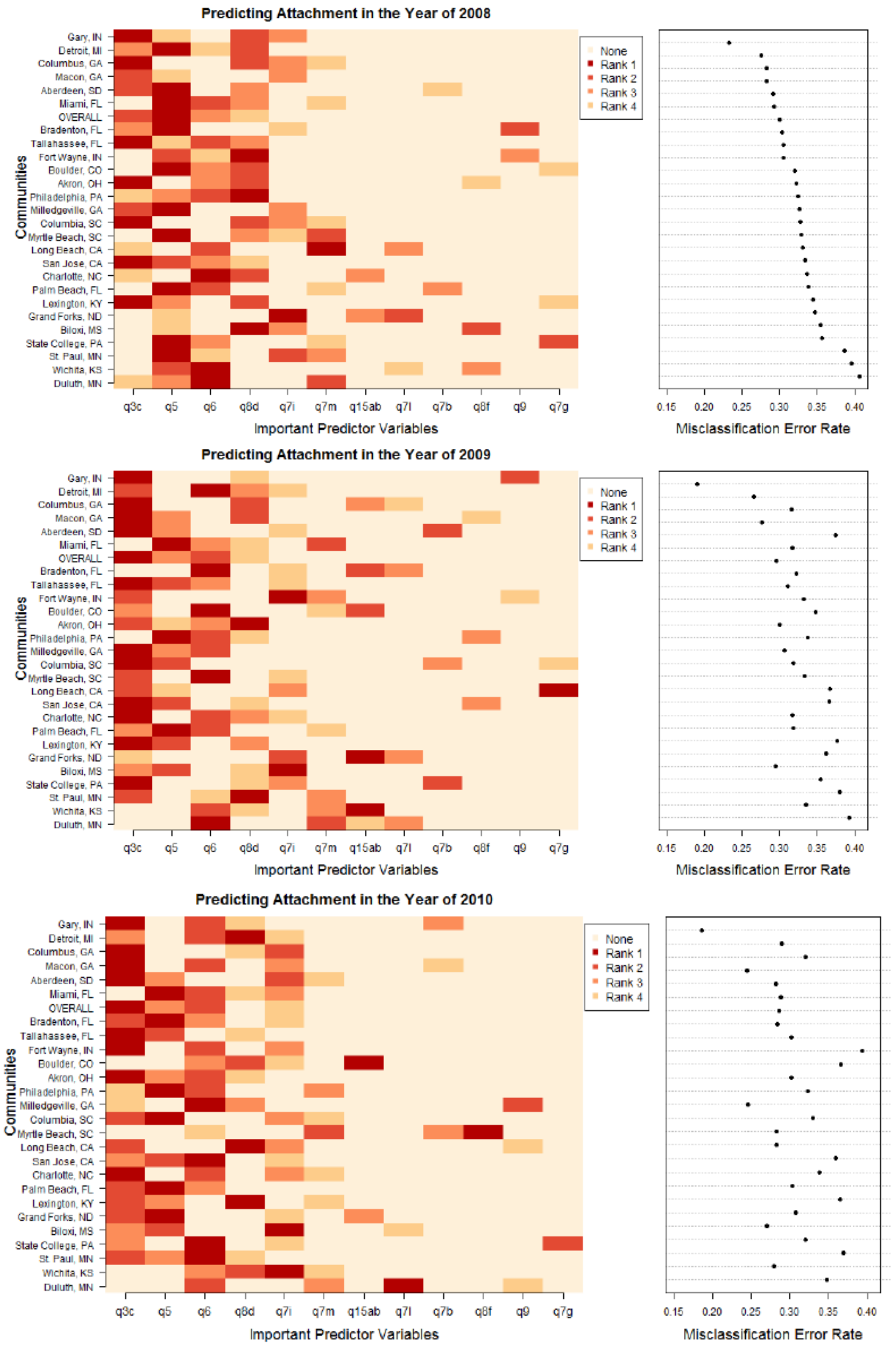


Fig. 3: Heatmaps of the random forests results (left) revealing the ranking of the important predictor variables in predicting attachment among all communities (labeled OVERALL) and in each community in each year. Variables (top 4 most important in each of the 81 models) are sorted from most frequent to least frequent occurrence in all communities and in each community in the three years. Communities are sorted according to the smallest misclassification error rate to the largest misclassification error rate in 2008. The misclassification error rates of the random forests results for each model are shown on the right. Note that the heatmap is truncated at 12 variables (out of the 18 variables that were determined to be important in at least one of the communities). See Table 4 in Appendix B for the meaning of each variable.

Results: Gary, IN

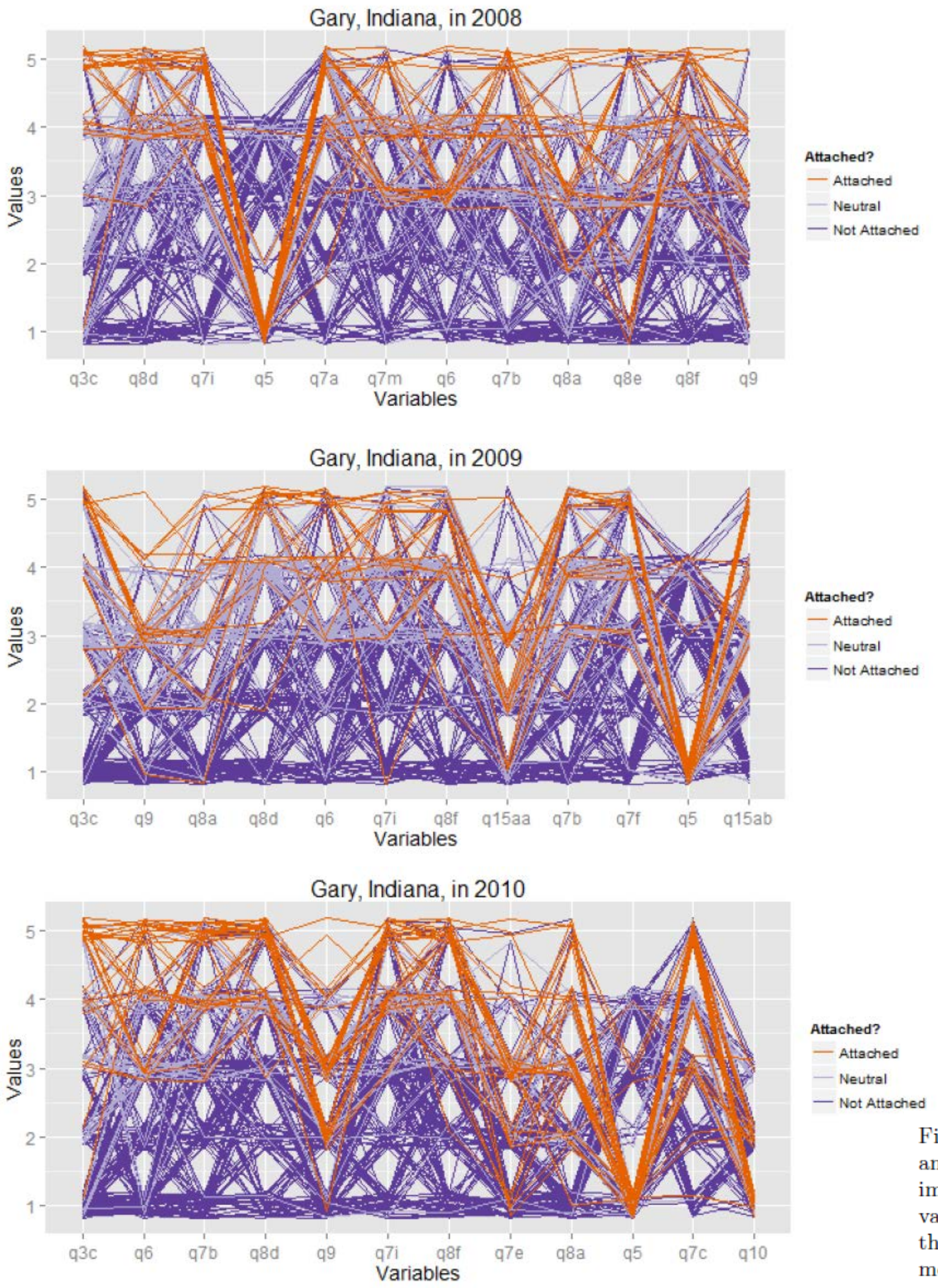


Fig. 6: Parallel coordinate plots for Gary, Indiana, for the years 2008, 2009, and 2010. Variables are ordered from most important (on the left) to the least important (on the right) as determined by the random forests algorithm. The values on the y-axis range from the minimum value among all the variables to the maximum value among all the variables. See Table 4 in Appendix B for the meaning of each variable.

Results: Archetypal Analysis

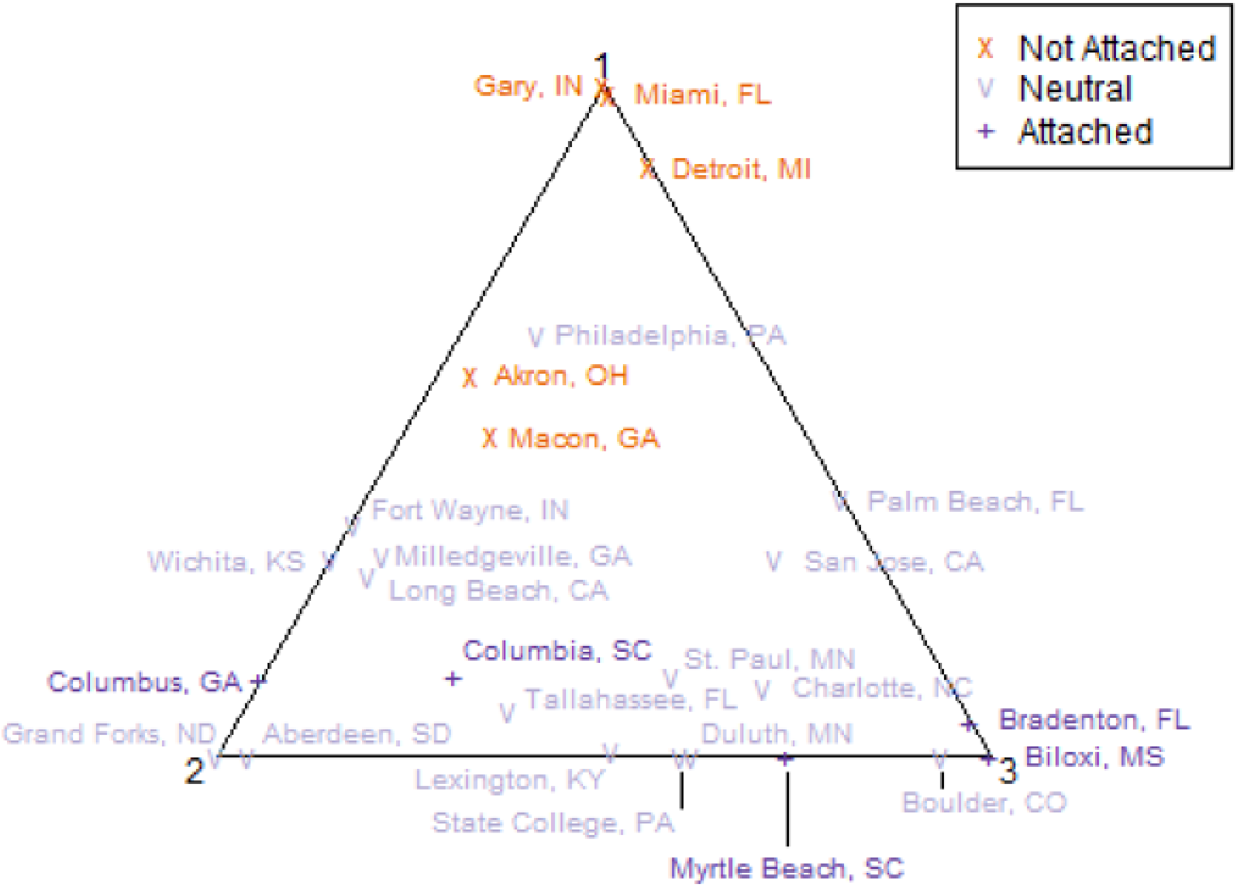


Fig. 7: Graphical representation of the three archetype solution for the year 2008. The three points labeled 1, 2, and 3 are the archetypes. Communities are colored by the dominating group according to the three attachment status levels.

Results: Breakdown of Variable q3c

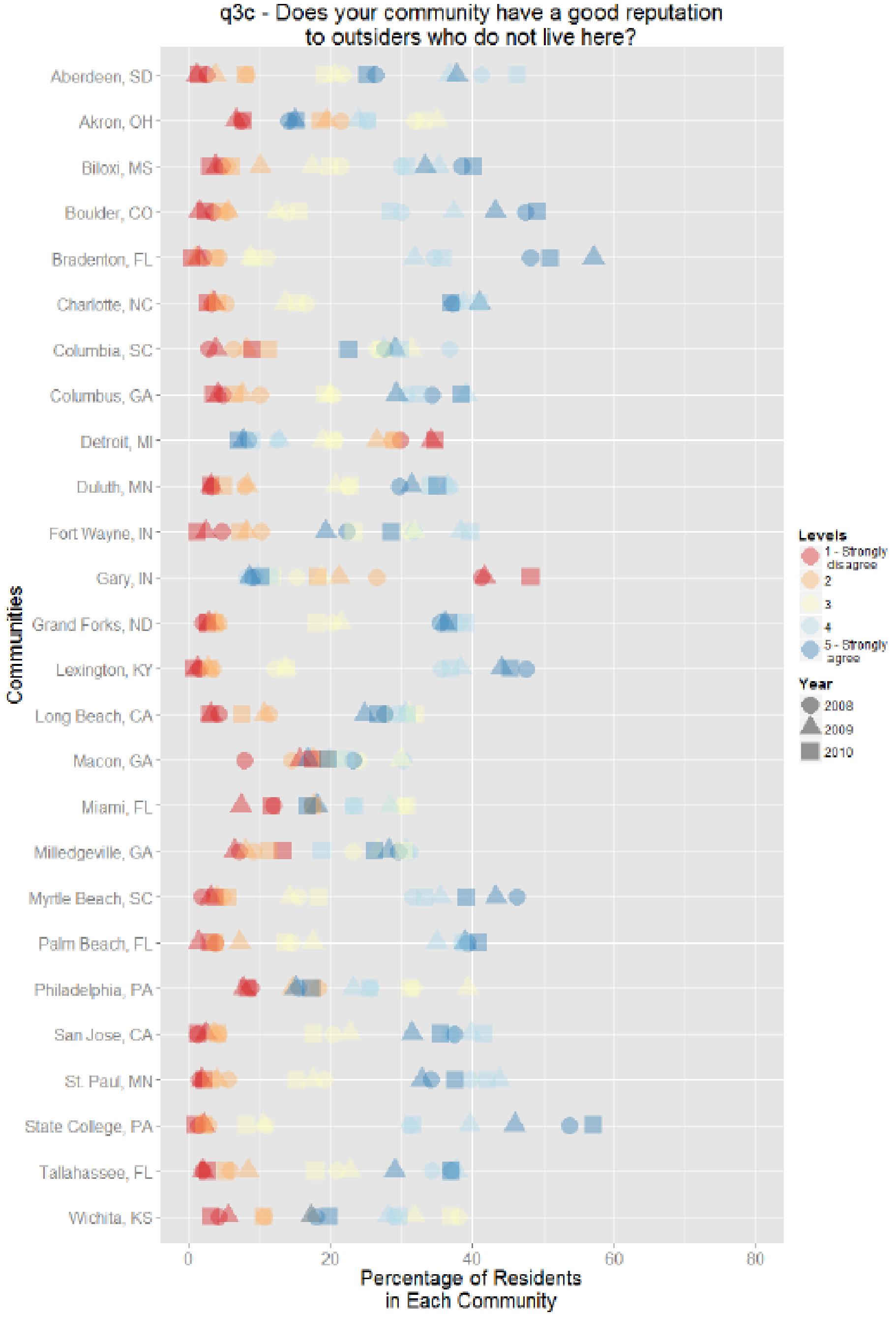


Fig. 9: Dot chart of the most important predictor variable (q3c). Points represent the percentage of people who responded in the community in that year.

Results: Breakdown of Variable q5

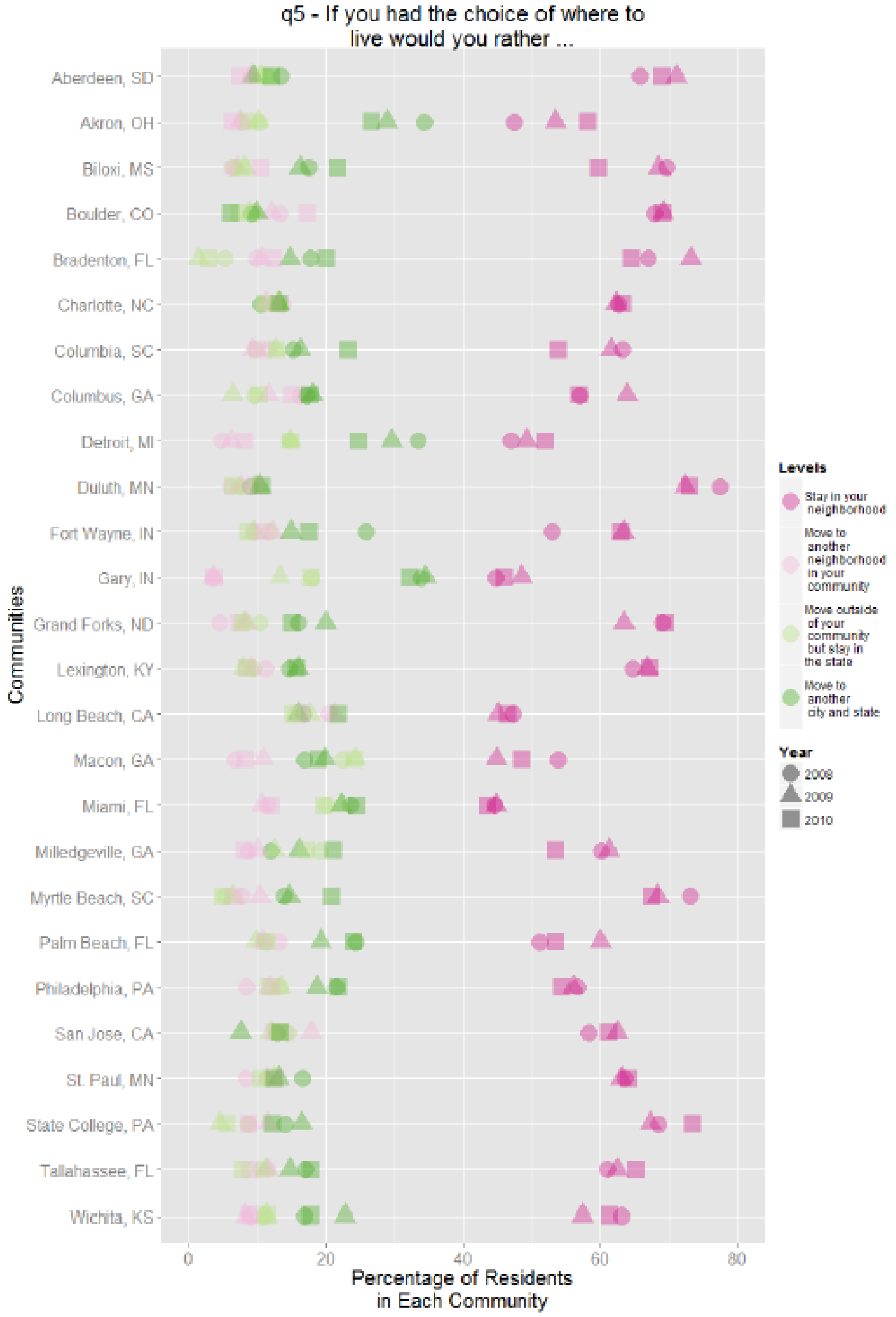


Fig. 10: Dot chart of the second most important predictor variable (q5). Points represent the percentage of people who responded in the community in that year.

Results: Breakdown of Variable q6

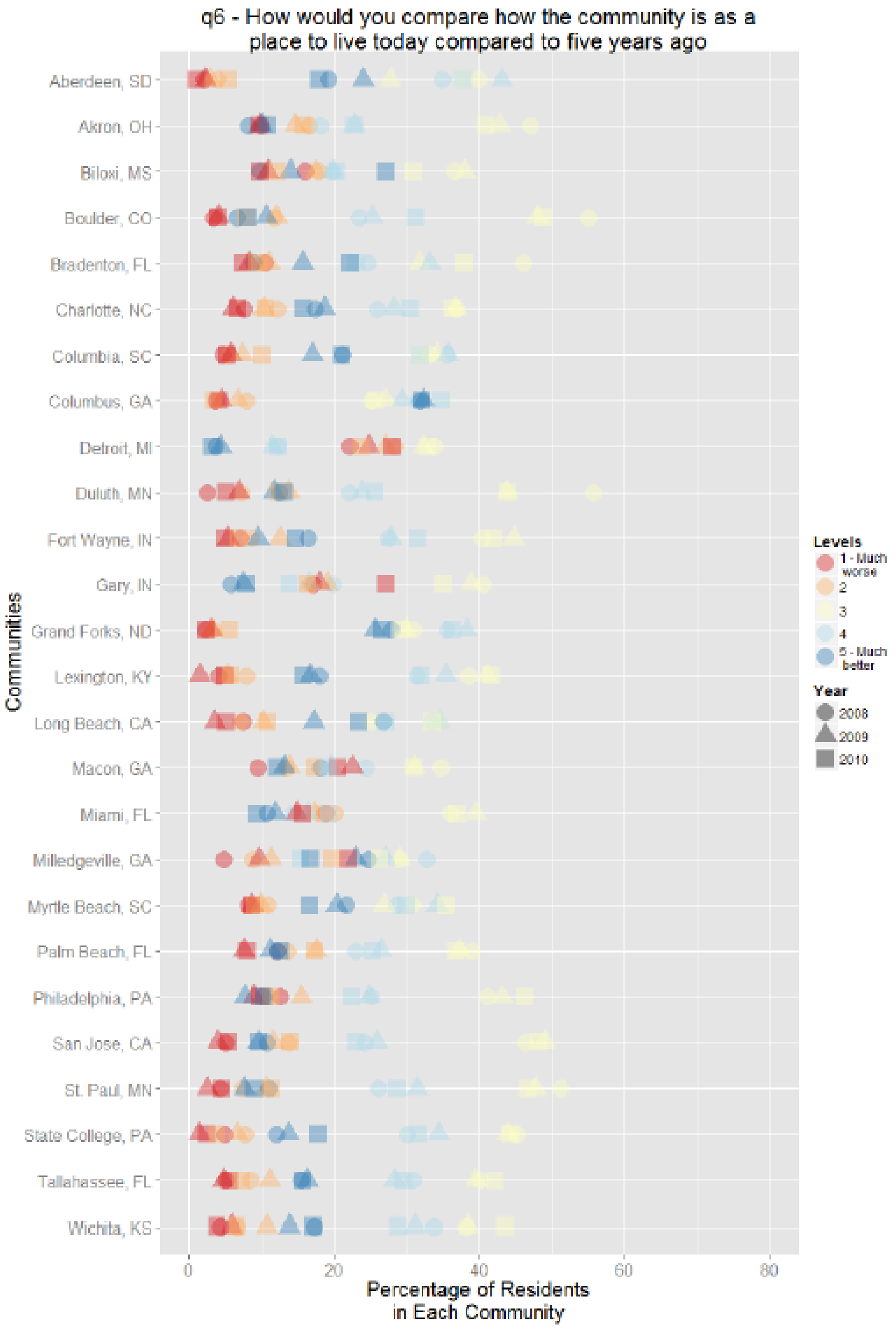


Fig. 11: Dot chart of the third most important predictor variable (q6). Points represent the percentage of people who responded in the community in that year.

Conclusion (1)

- Mainly three variables are important in determining attachment status:
  - q3c (the community has a good reputation to outsiders or visitors who do not live here)
  - q5 (if you had the choice of where to live would you rather ...)
  - q6 (how would you compare how the community is as a place to live today compared to 5 years ago?)

Conclusion (2)

- Some communities are rather unusual (when compared to the other communities) with respect to some of the predictor variables
- Overall, people who have positive things to say about their community were also attached to their community and wanted to stay within their neighborhood

Case Study 3: Forecasting of “Snow Water Equivalent” Measurements

Published as:

Odei, J. B., Symanzik, J., Hooten, M. B. (2014): A Bayesian Hierarchical Model For Forecasting Intermountain Snow Dynamics, *Environmetrics*, 25(5): 324-340.

Background (1)

- The Intermountain region of the Western United States comprises of a variety of ecological and economic systems
- Snowpack – accounts for 50 to 70% of the annual precipitation in the Intermountain regions (Serreze et al., 1999)
- Over 75% of its water resources results from snowmelt water
- Multi-year droughts in the Southwest have severely affected supplies according to a report from the National Climatic Data Center
- These droughts are among major natural risks this region’s residents and ecosystems are facing

Background (2)

- Difficulties associated with accurately determining the time of maximum accumulation present a problem for snowmelt runoff forecasters
- Various approaches to estimating snow pack characteristics differ in spatial scale, reliability, and accuracy
- The U.S. Department of Agriculture (USDA) operates the Snow Telemetry (SNOTEL) system, which is a network of remote, automatic, monitoring stations that yield online daily measurements of SWE, precipitation, temperature and, more recently, snow depth
- To forecast water resources, the National Weather Service (NWS) maintains a set of conceptual, continuous, hydrologic simulation models used to generate extended streamflow outlooks, and flood forecasts

Problems Associated with Current Snow Models

- The empirical statistical models developed, lack some of the characteristics required for evaluating large mammal dynamics, such as:
  - **(1.)** scaling up to large study areas
  - **(2.)** incorporating temporal dynamics, deterministic results, or an objective, validated basis
- General Circulation Models (GCMs) – Unable to adequately capture snow-related atmospheric processes in mountainous areas

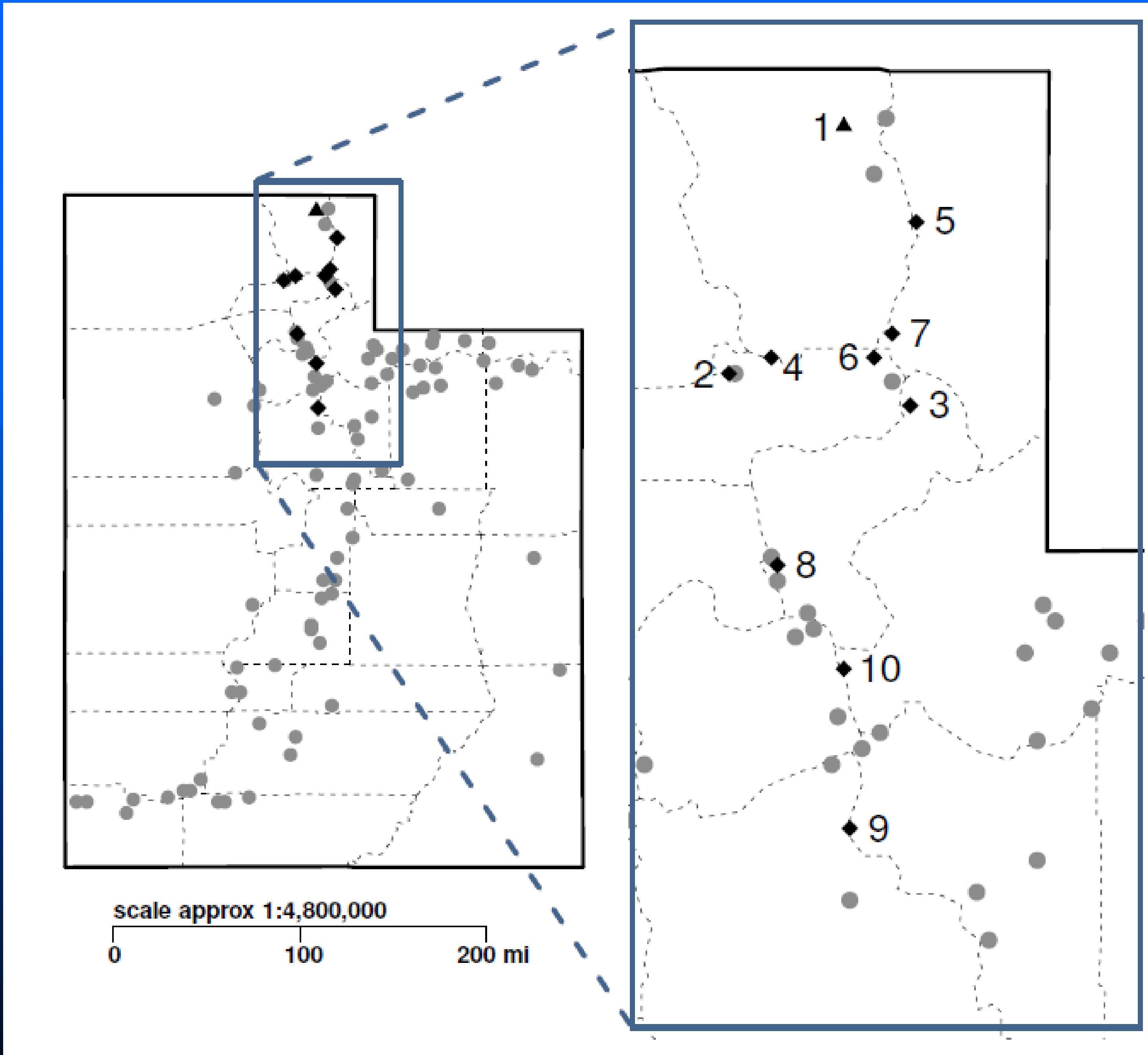
A Bayesian Hierarchical Model (1)

- Addressing issues of modeling SWE – utilization of statistically based snow models that rely heavily on observational data
- A general spatio-temporal statistical model was introduced in Odei et al., (2009)
  - A simplified version initially treats the SNOTEL sites independently
  - Development of a hierarchical statistical model for SWE data using a Bayesian approach

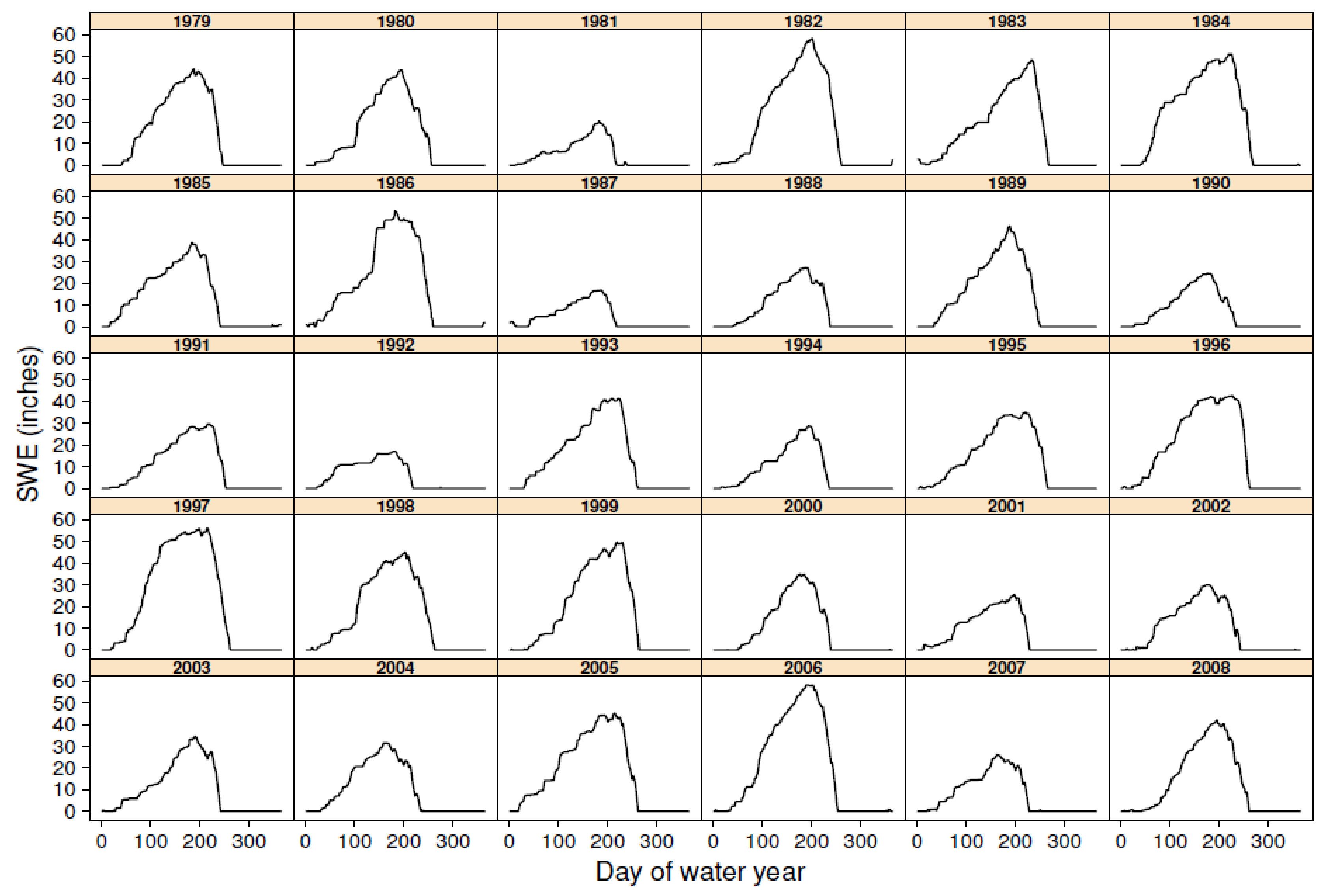
A Bayesian Hierarchical Model (2)

- For site s, predict SWE for all days $\geq$ t of the water-year
- Data used:
  - Averages of SWE measured for each day in the water-year at site s for the past T water-years
  - Temporal (daily) SWE correlation
  - SWE at site s up to day t-1 in current water-year

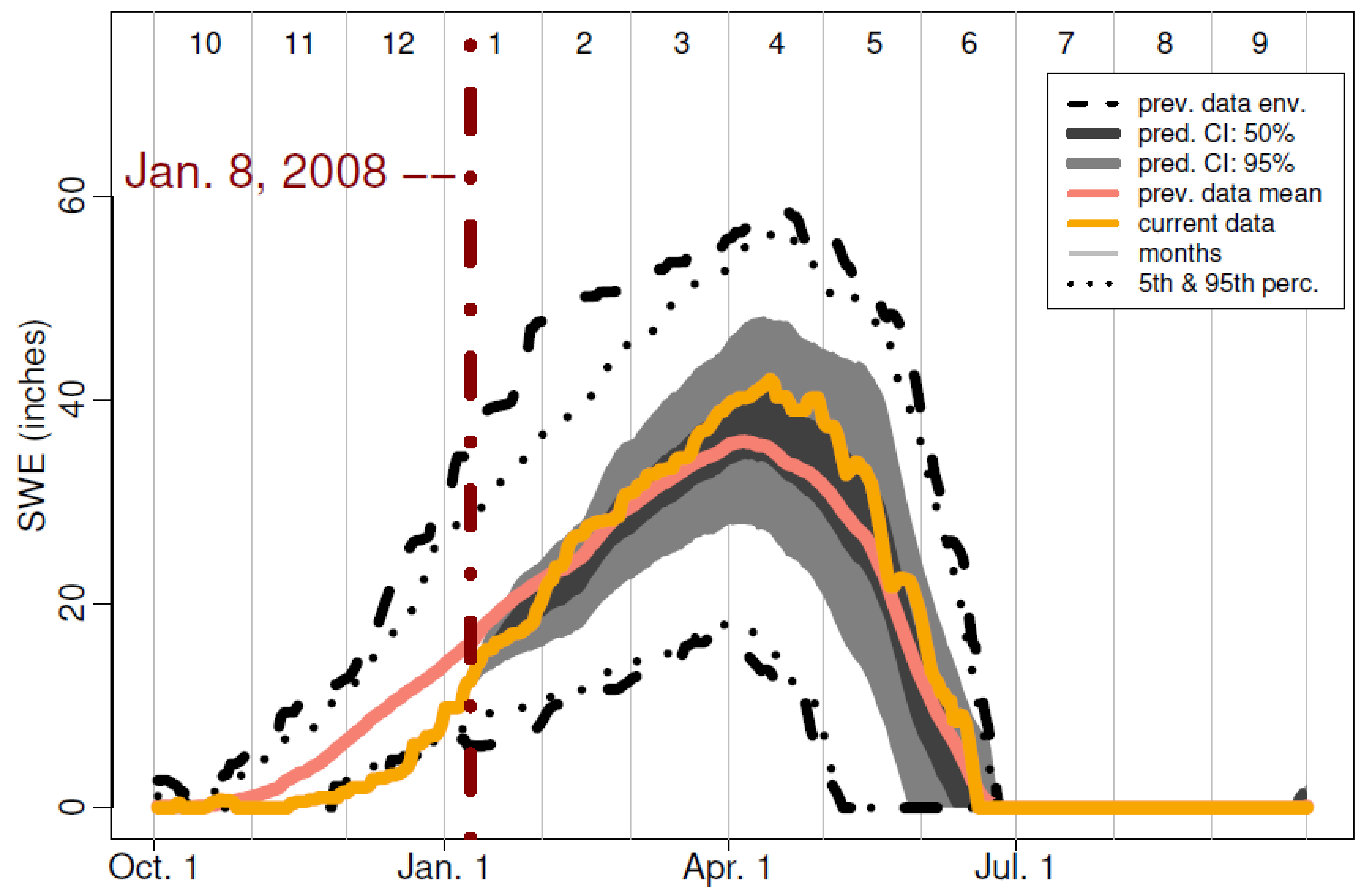
Utah SNOTEL Sites



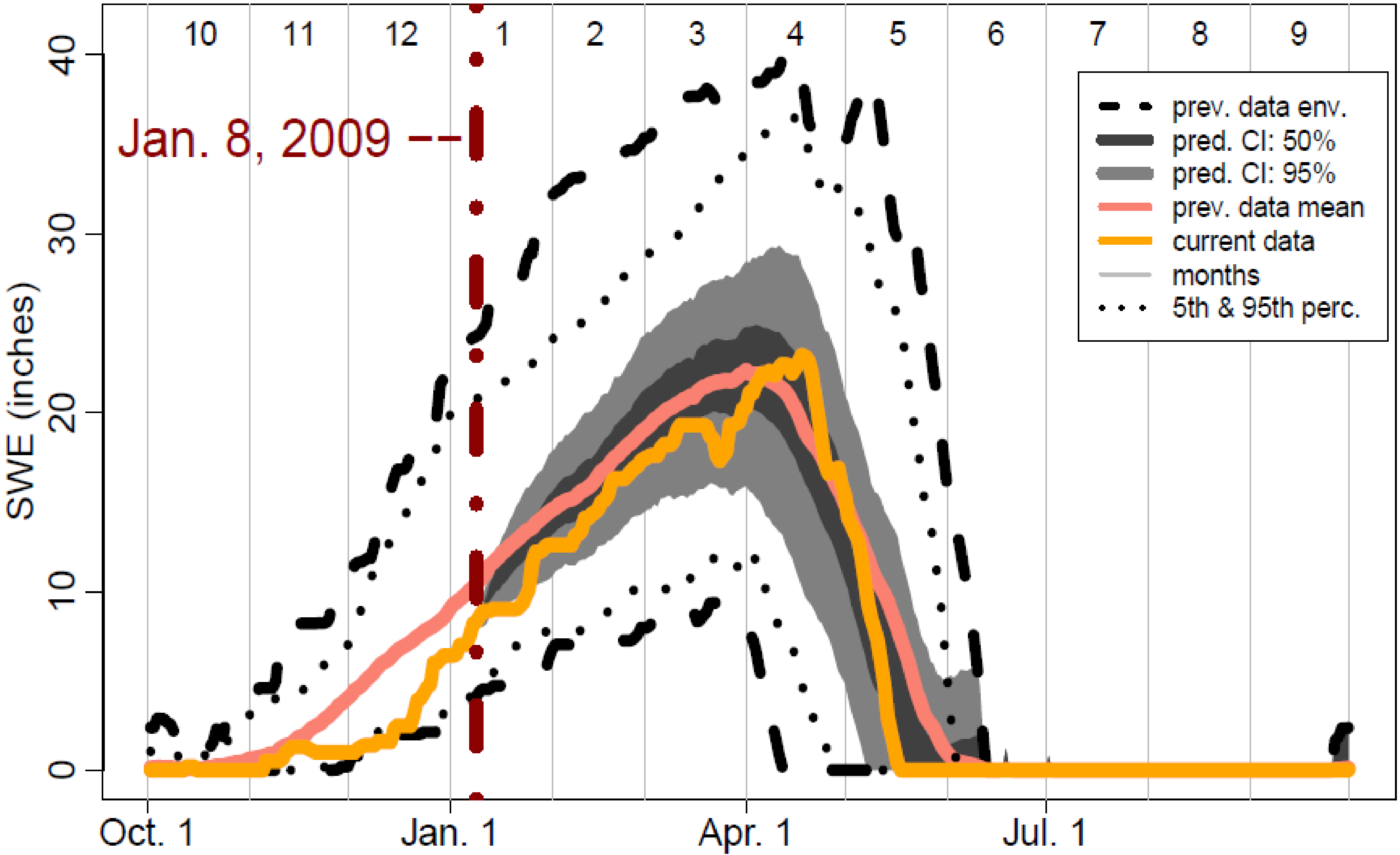
Tony Grove SNOTEL Site, Utah – SWE Measurements



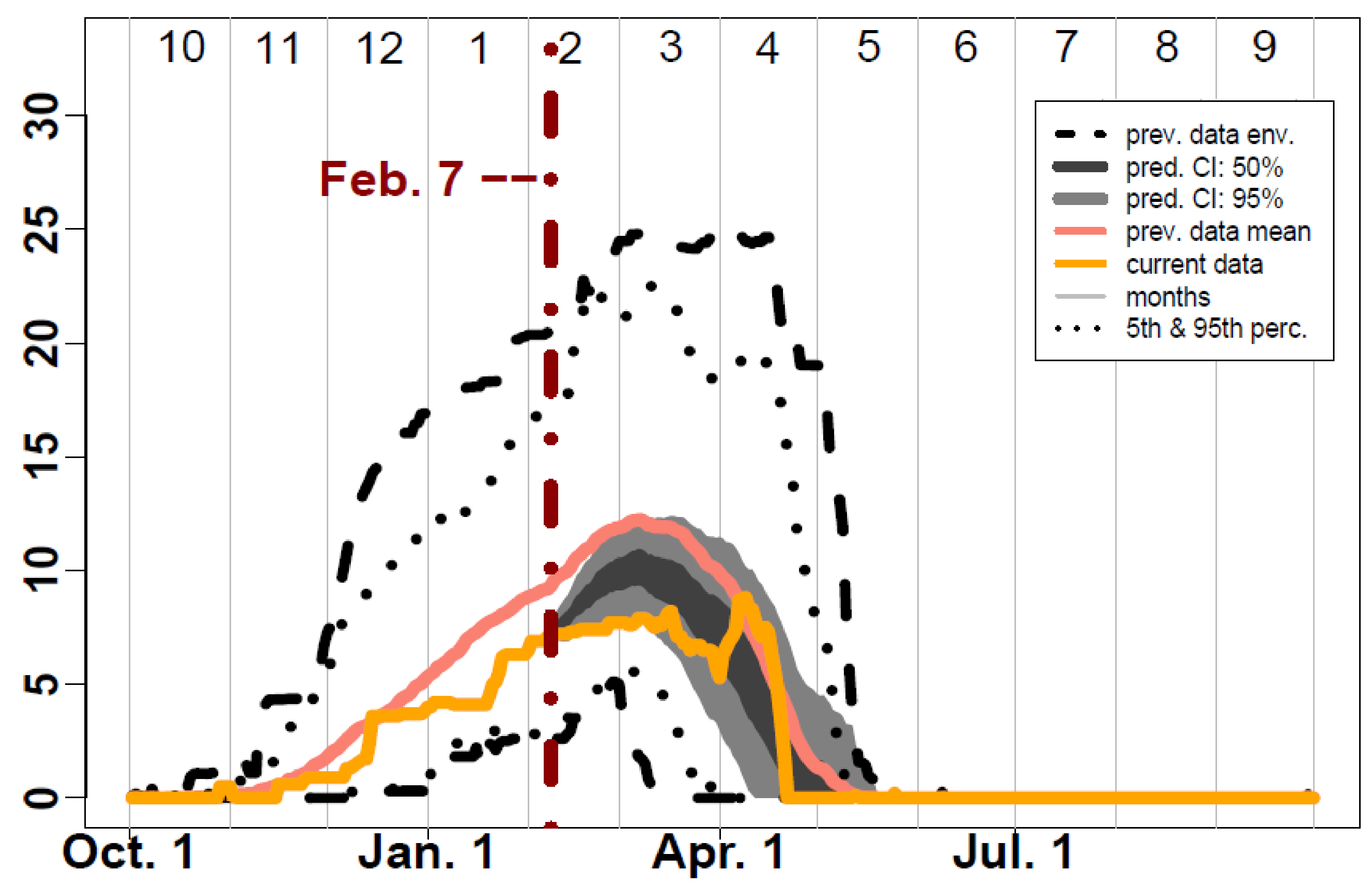
Result: Tony Grove SNOTEL Site, Utah – 2008 Water-Year



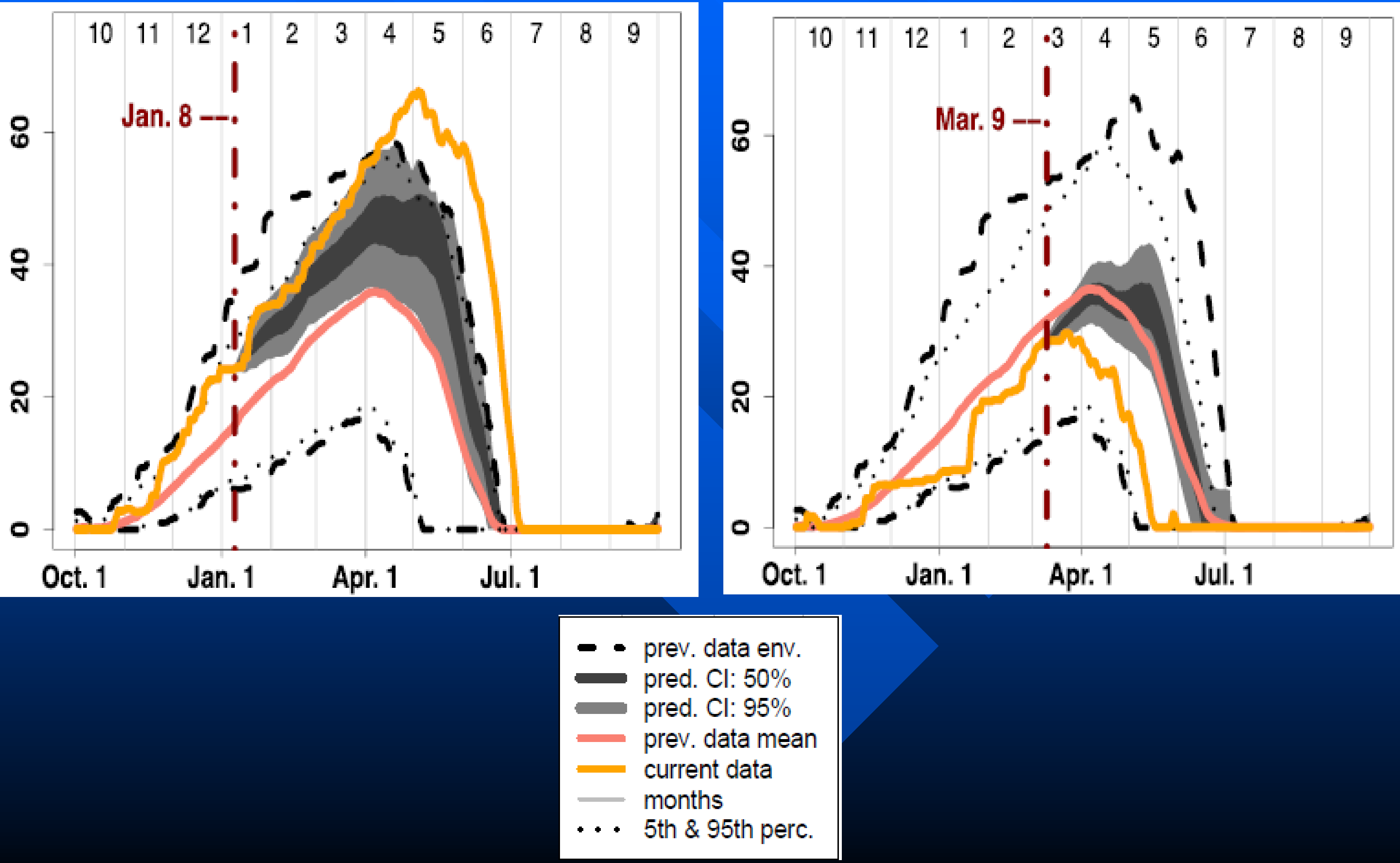
Result: Horse Ridge SNOTEL Site, Utah – 2009 Water-Year



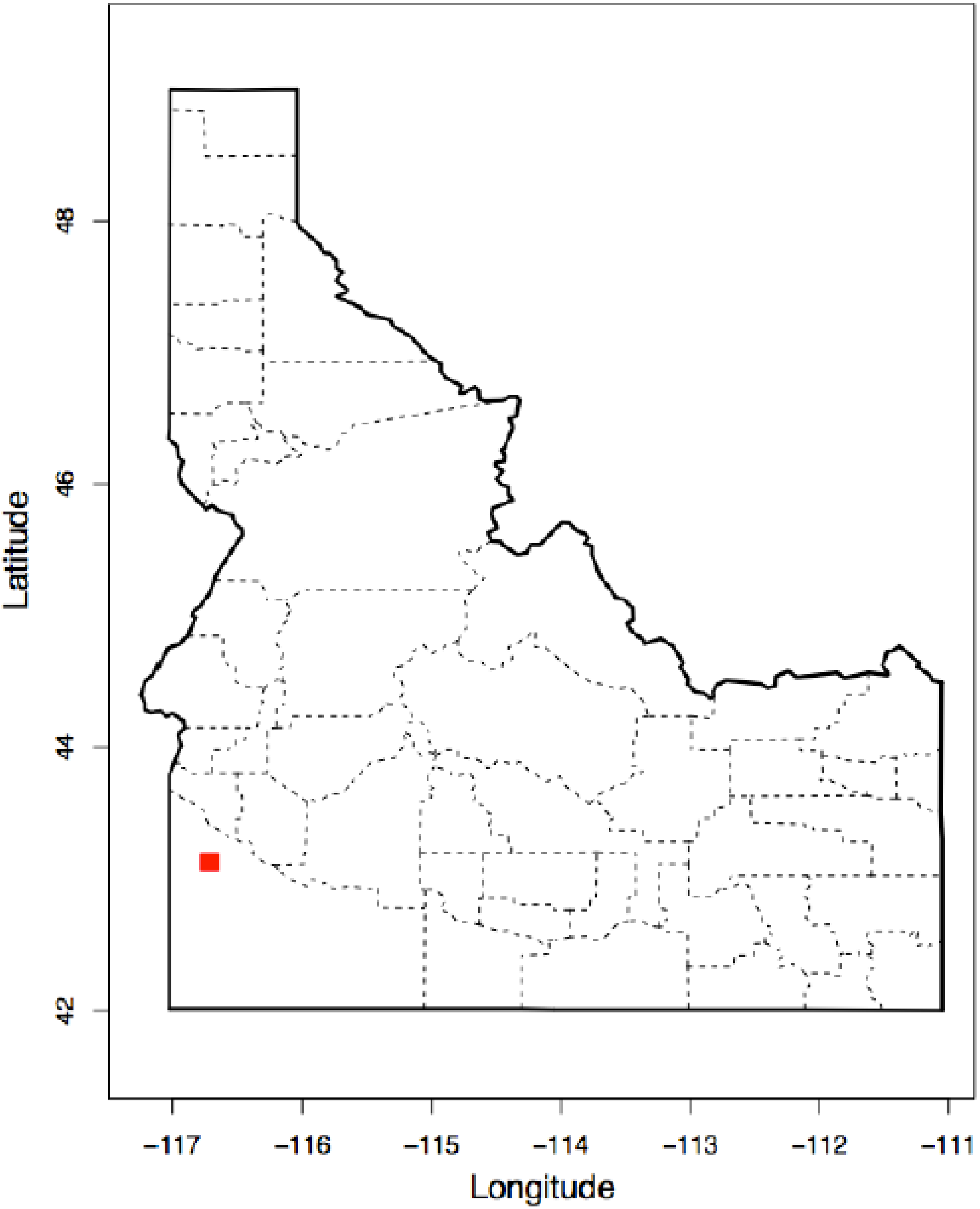
Result: Little Bear SNOTEL Site, Utah – 2010 Water-Year



Result: Tony Grove Prediction Failures

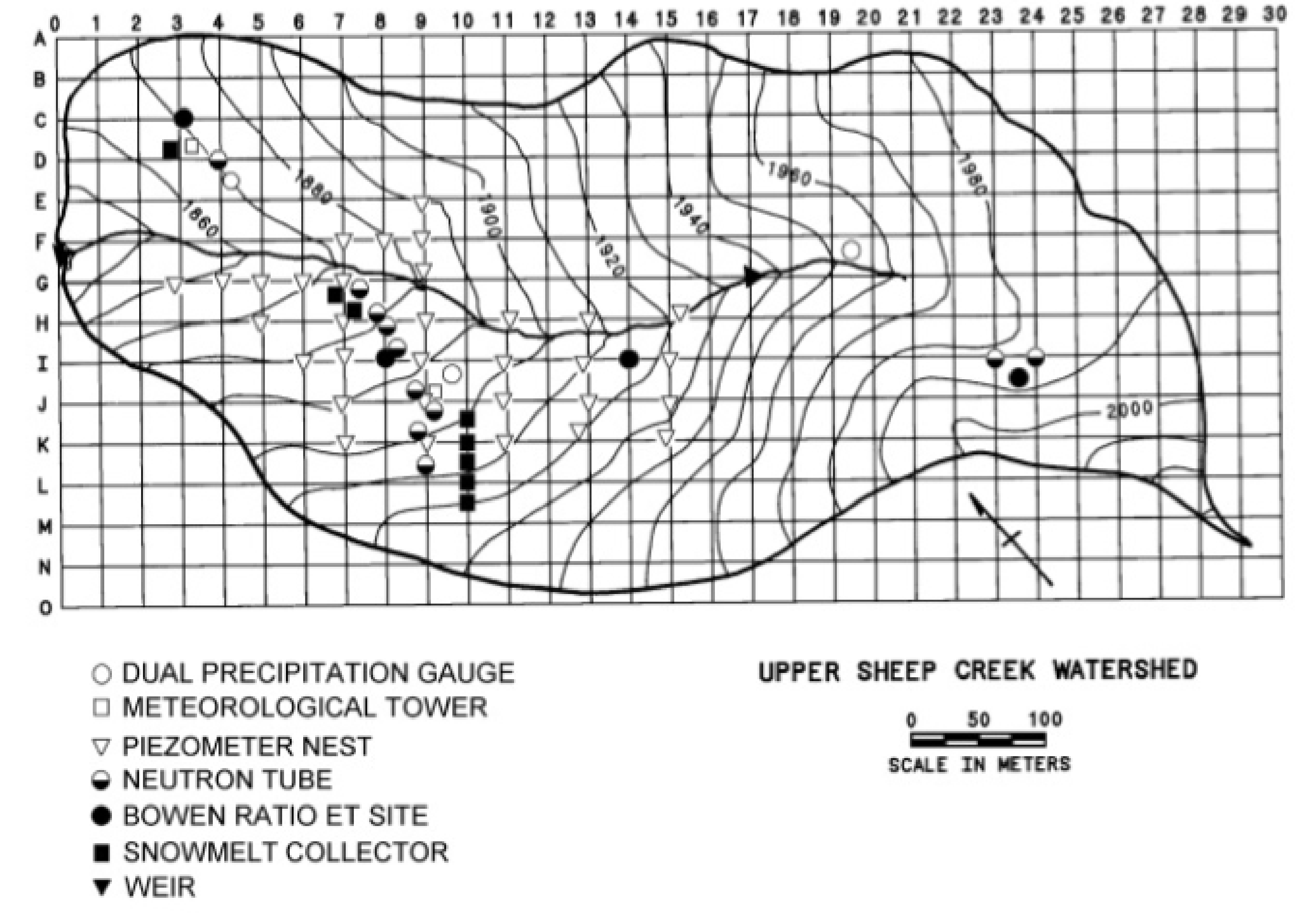


Upper Sheep Creek Watershed Data, Idaho (1)



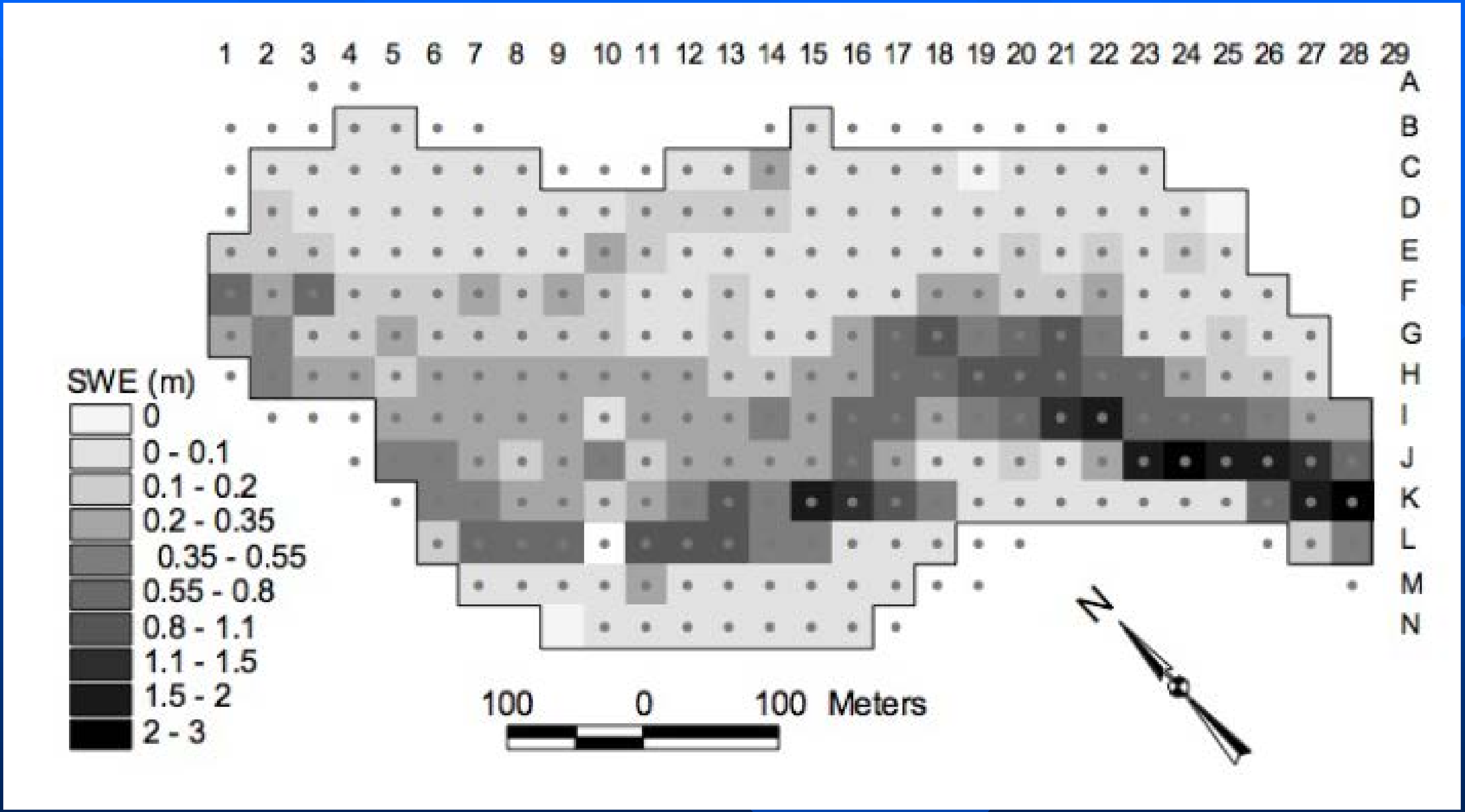
Additional Example from Odei (2014)

Upper Sheep Creek Watershed Data, Idaho (2)



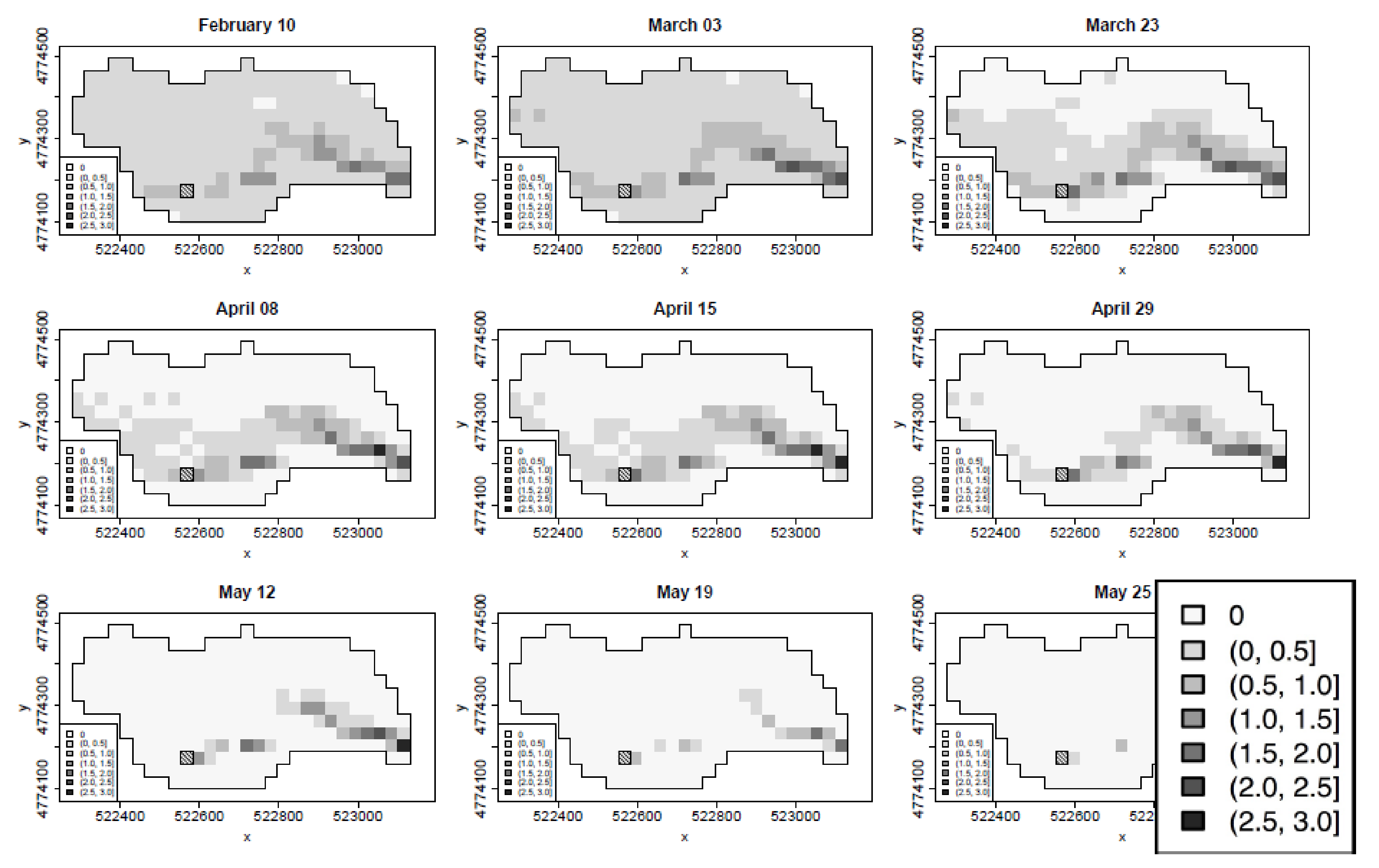
Topography and instrument locations within the Upper Sheep Creek Watershed (Previously published as Figure 1 in Flerchinger and Cooley (2000))

Upper Sheep Creek Watershed Data, Idaho (3)

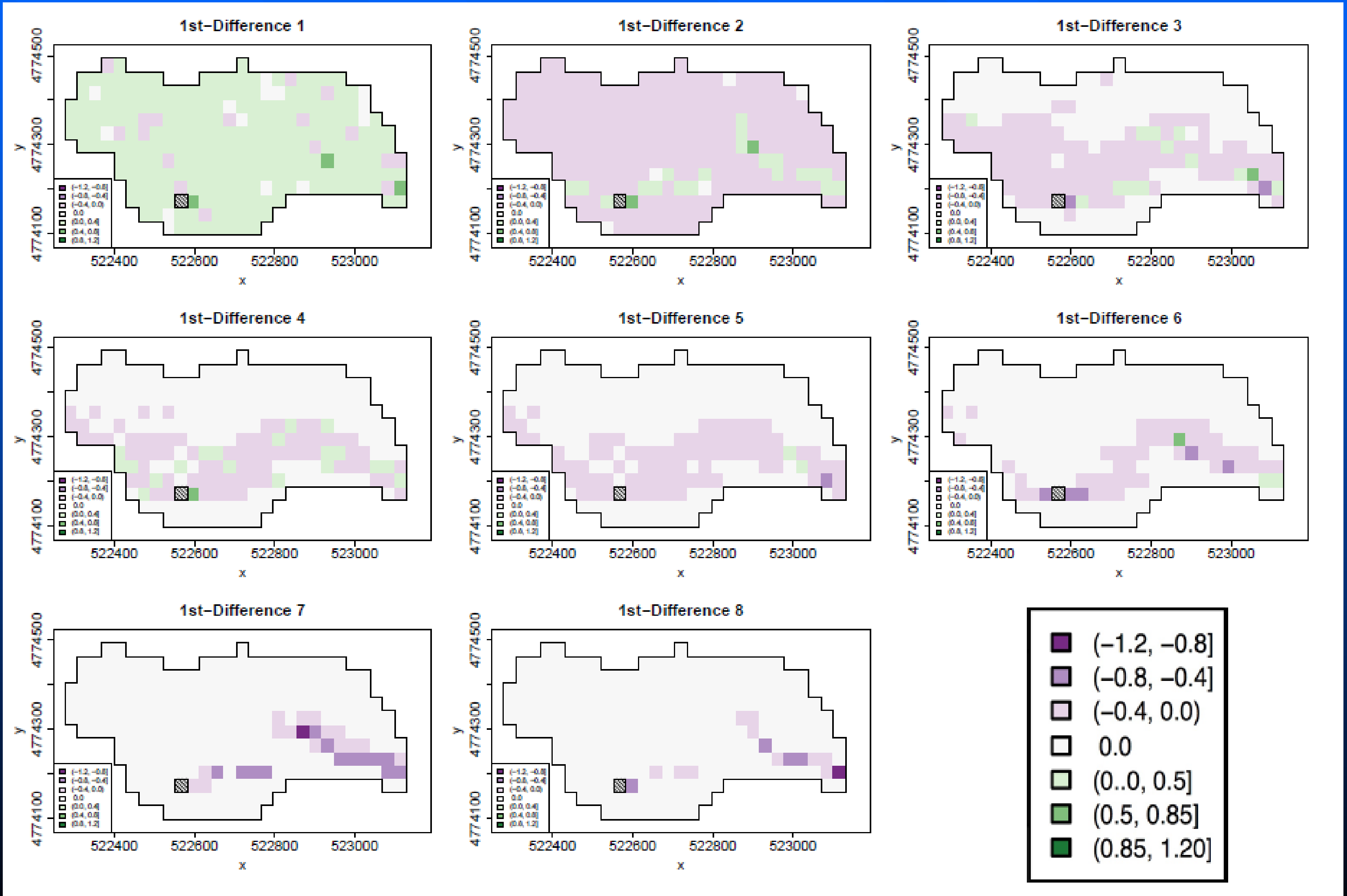


SWE measured in Upper Sheep Creek March 3, 1993. The dots represent locations of the grid stations where measurements of SWE were taken. No grid stations are available at points 9N and 25D and point L10 was not measured (Previously published as Figure 2 in Luce and Tarboton (2004))

Upper Sheep Creek – Temporal Micromaps



Upper Sheep Creek – Temporal Comparative Micromaps



Conclusions

- Visual approach helps to effectively display the SWE forecasts from a complex statistical model
- Visual approach helps to effectively assess changes in temporal SWE data

Further Reading

- Flerchinger, G. N., Cooley, K. R. (2000). A Ten-Year Water Balance of a Mountainous Semi-Arid Watershed. *Journal of Hydrology*, **237**, 86–99.
- Luce, C. H., Tarboton, D. G. (2004). The Application of Depletion Curves for Parameterization of Subgrid Variability of Snow. *Hydrological Processes*, **18**, 1409–1422.
- Odei, J. B. (2014). Statistical Modeling, Exploration, and Visualization of Snow Water Equivalent Data, Dissertation. http://digitalcommons.usu.edu/etd/3871
- Odei, J. B., Hooten, M. B., Jin, J. (2009). Inter-Annual Modeling and Seasonal Forecasting of Intermountain Snowpack Dynamics, in: *2009 JSM Proceedings*, American Statistical Association, Alexandria, Virginia, pp. 870–878. (CD).
- Serreze, C. M., Clark, P., Amstrong, R. L., McGinnis, D. A., Pulwarty, R. S. (1999). Characteristics of Western U.S. Snowpack from Snotel Data, *Water Resources Research*, **35**, 2145–2160.

Case Study 4: Exploratory Graphics for Functional Actigraphy Data in Sleep Medicine

Published as:

Symanzik, J., Shannon, W. (2008): Exploratory Graphics for Functional Actigraphy Data, *JSM Proceedings*, American Statistical Association, CD.

Ding, J., Symanzik, J., Sharif, A., Wang, J., Duntley, S., Shannon, W. D. (2011): Powerful Actigraphy Data Through Functional Representation, *Chance*, 24(3): 30-36.

Sharif, A., Symanzik, J. (2012): Graphical Representation of Clustered Functional Actigraphy Data, 2012 *JSM Proceedings*, American Statistical Association, Alexandria, Virginia, CD.

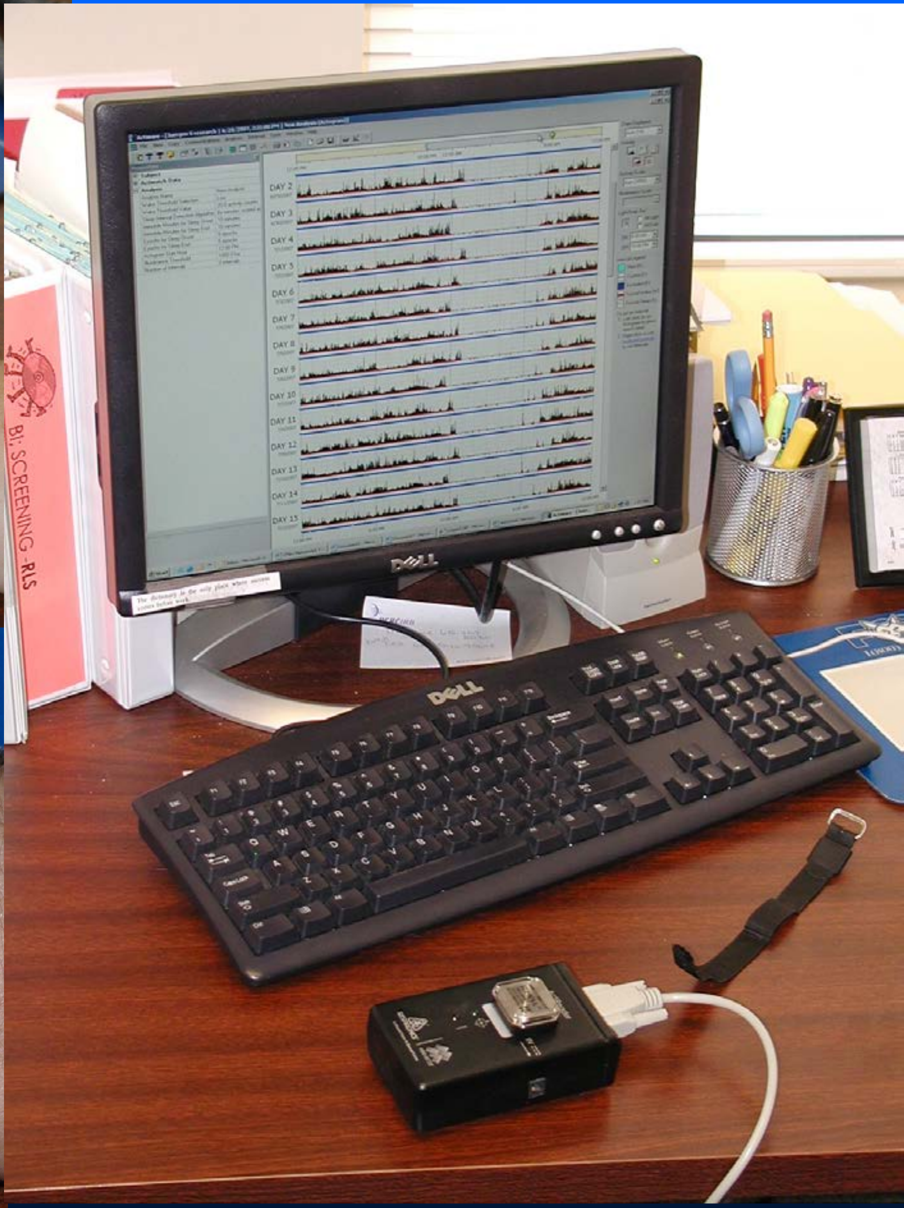
Sharif, A., Symanzik, J. (2013): ActiVis, an R Package for the Visualization of Functional Actigraphy Data (and Beyond), In: Cho, S.-H. (Ed.), *Proceedings of Joint Meeting of the IASC Satellite Conference and the 8th Conference of the Asian Regional Section of the IASC*, Asian Regional Section of the IASC, 145-150.

Background (1)

- Actigraphy: emerging technology for measuring a patient’s activity level continuously over time
- Actigraph: watch-like device (attached to the wrist or a leg) that uses an accelerometer to measures (human) movements (every minute or more often)

• Respiratory
• Accessibility
• Rehab
AP/Bi-Level
• Bronchodilators
• DME

Actiwatch®
Ctrl


DAY 2
DAY 3
DAY 4
DAY 5
DAY 6
DAY 7
DAY 8
DAY 9
DAY 10
DAY 11
DAY 12
DAY 13
DAY 14
DAY 15
DELL
BI; SCREENING -RLS

Background (2)

- Analysis of Human Actigraphy Data: Useful for detecting sleep, for assessing insomnia and restless leg syndrome, for tracking recovery after heart attacks, and as an assessment tool for overall status of HIV patients
- Actigraphy Data can be best described as functional data

Visualization of Functional Data

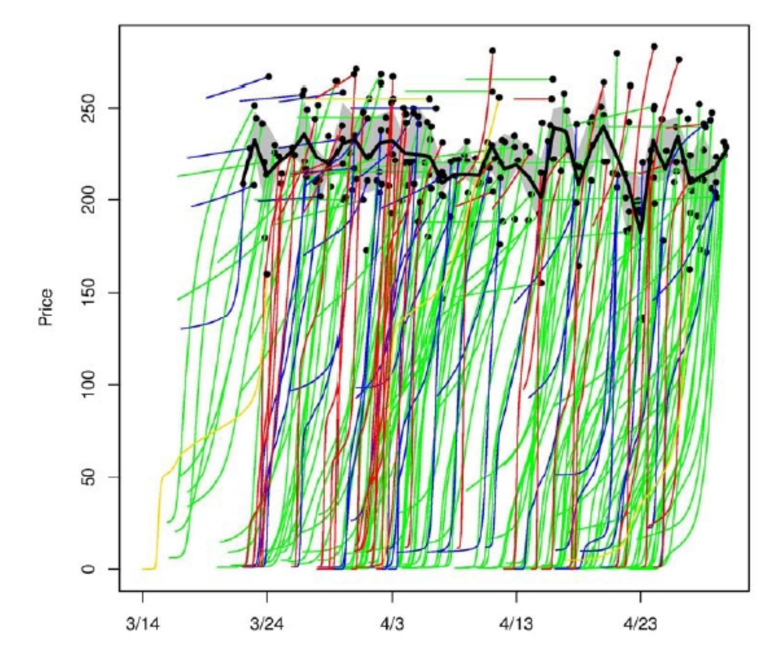
- Very limited ! A rare example is:

  Jank, W., Shmueli, G., Plaisant, C., Shneiderman, B. (2008): Visualizing Functional Data with an Application to eBay's Online Auctions, In: Chen, C., Härdle, W., Unwin, A. (Eds.), Handbook of Data Visualization, Springer, Berlin/Heidelberg, 873-898.

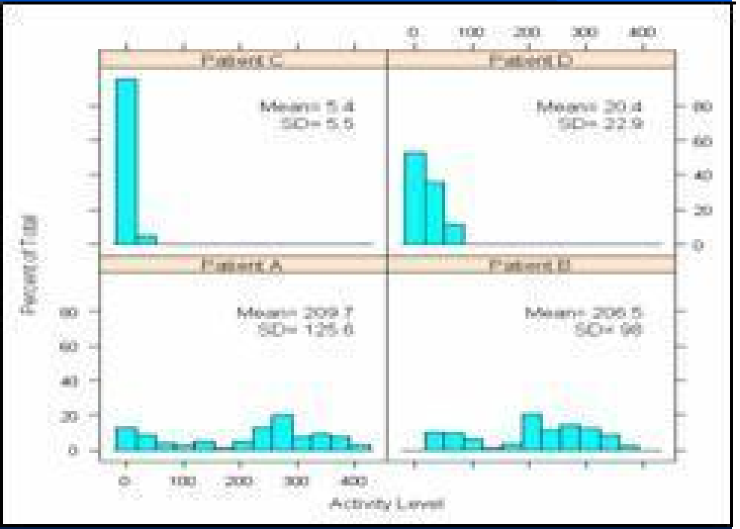
- Figure from

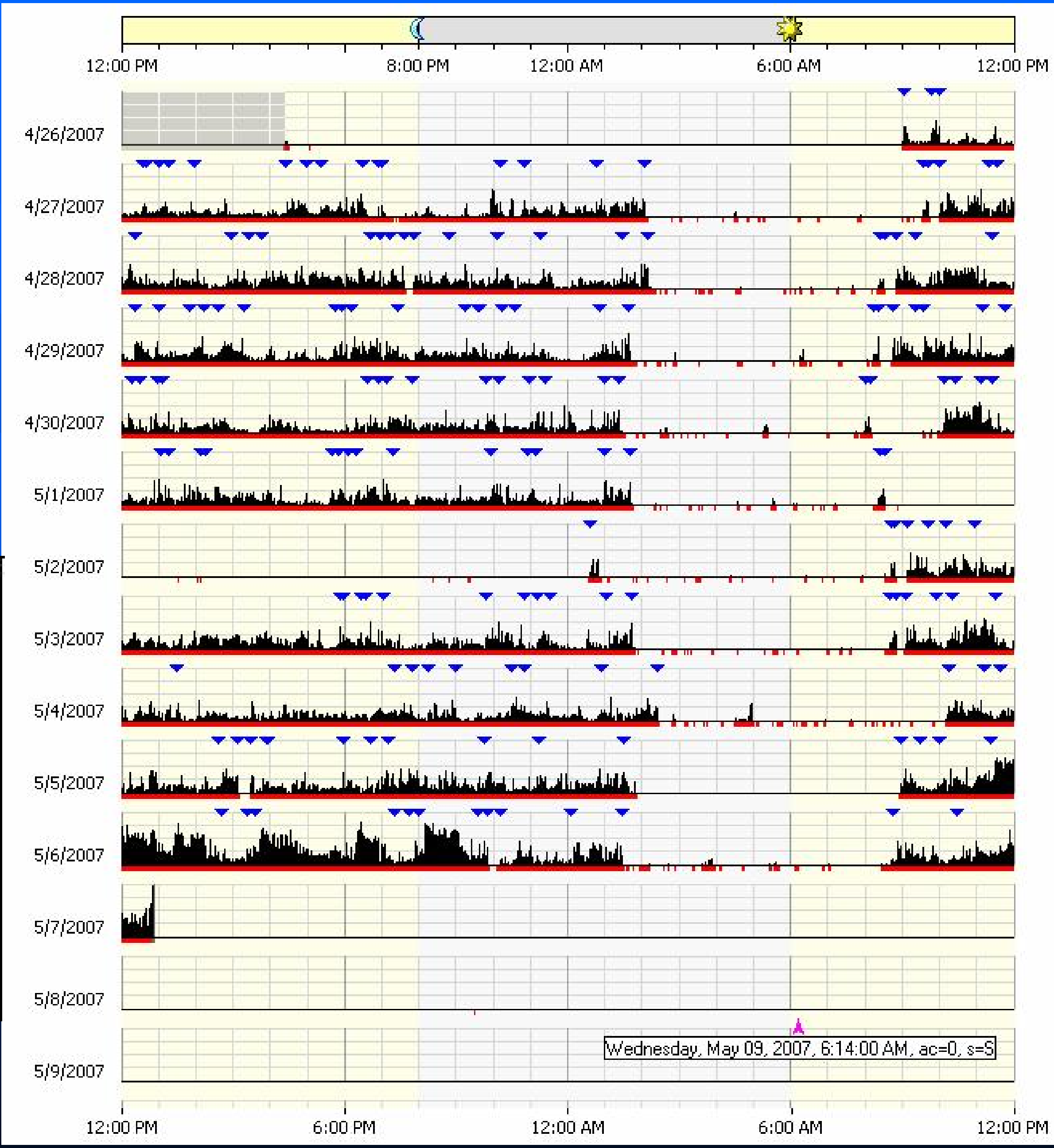
  http://www.smith.umd.edu/faculty/wjank/DIV-Berlin2006.pdf

  (page 30).



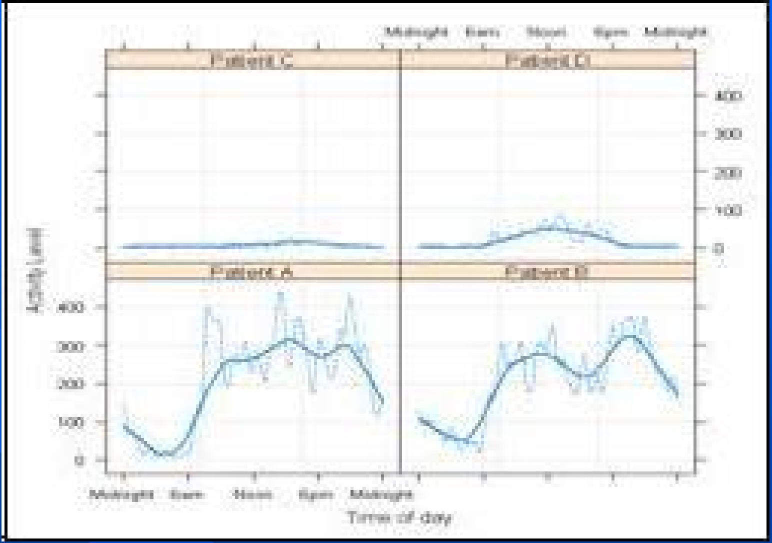
Current Visualization of Actigraphy Data

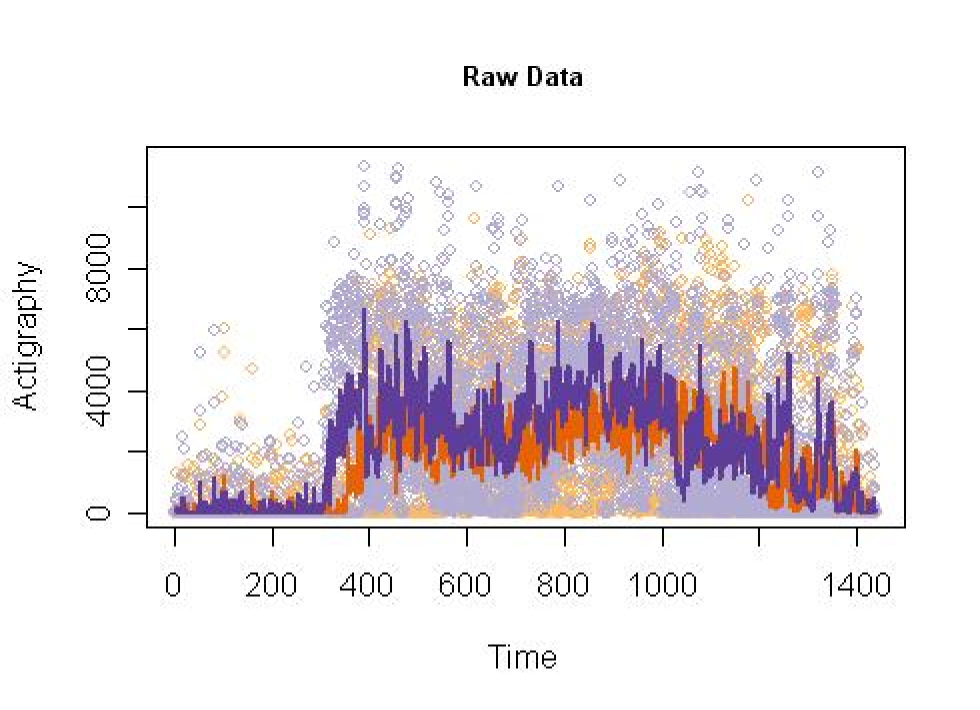


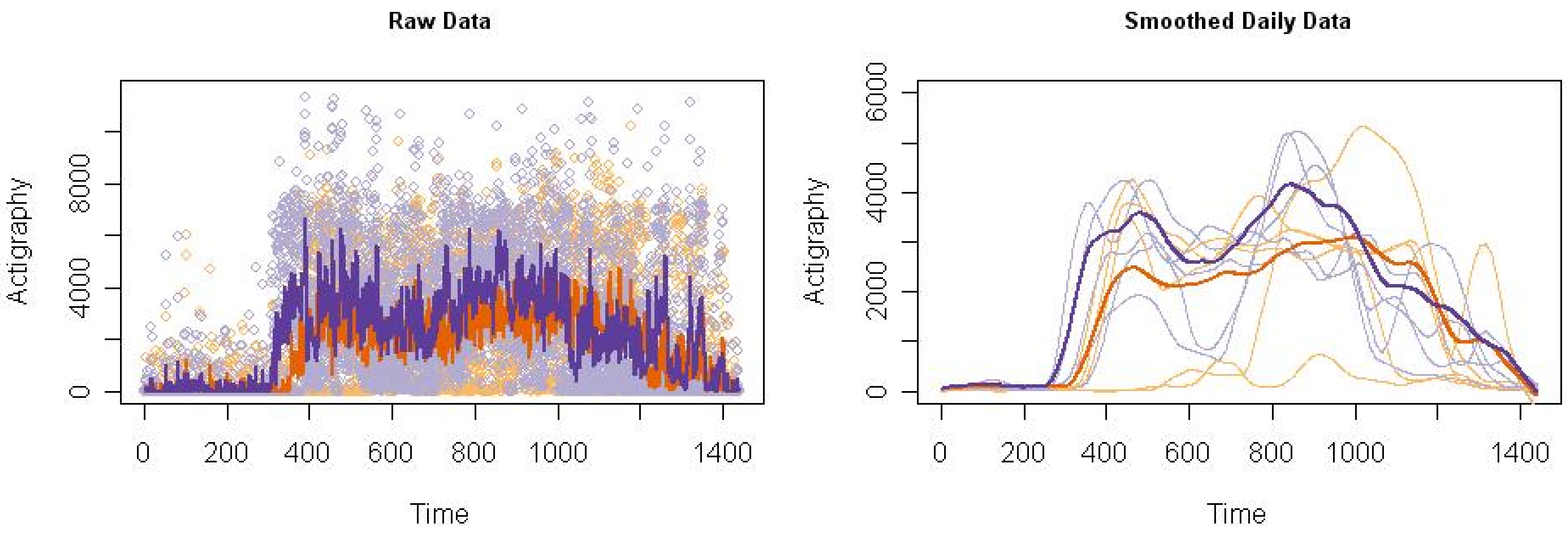


Suggested Future Visualization of Actigraphy Data (1)

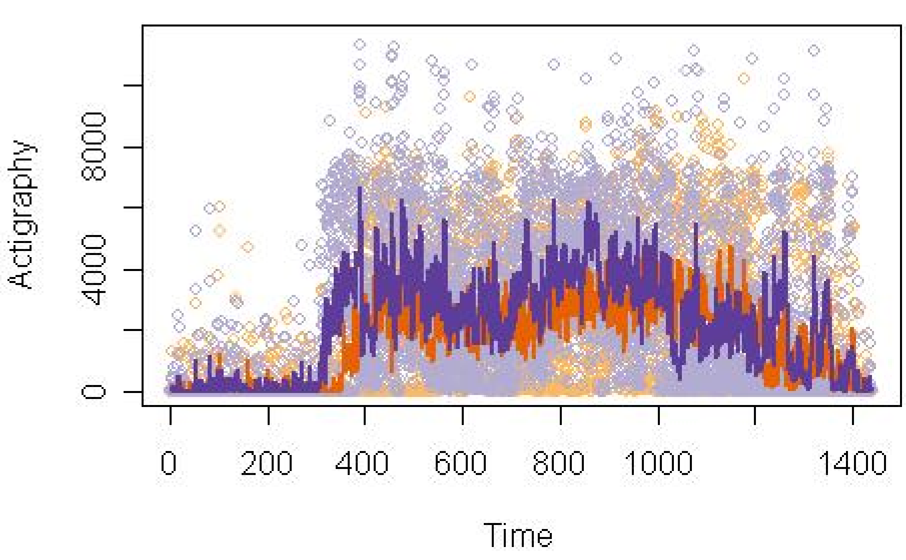
- Displays for
  - Raw data
  - Smoothed data
  - Averages etc.
  - Velocity (First Derivative)
  - Acceleration (Second Derivative)
  - Brushing & Linking
  - Cumulative Sums
- Example: 1 Subject
  - Orange: 5 Days at Baseline
  - Purple: 5 Days after 6 Months



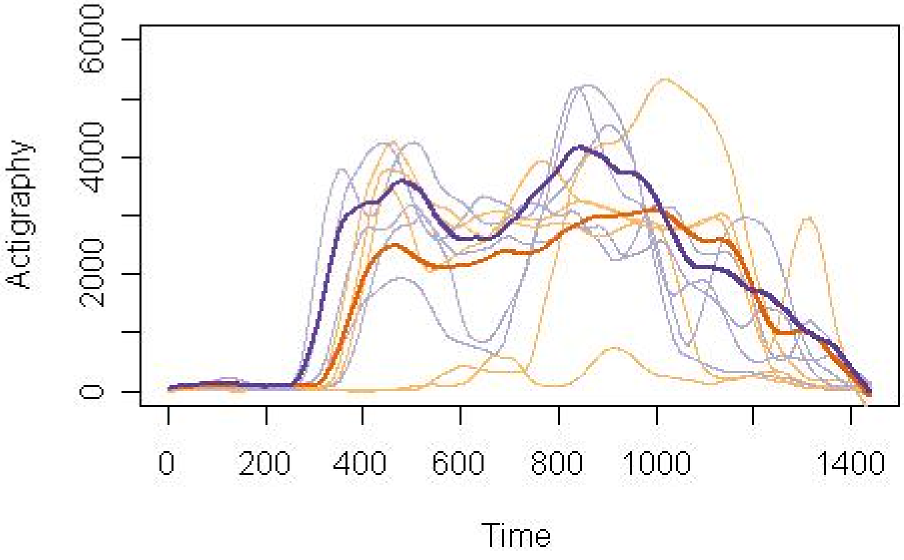

Raw Data
Actigraphy
0
4000
8000
0
200
400
600
800
1000
1400
Time


Raw Data
Actigraphy
0
4000
8000
0
200
400
600
800
1000
1400
Time
Smoothed Daily Data
Actigraphy
0
2000
4000
6000
0
200
400
600
800
1000
1400
Time

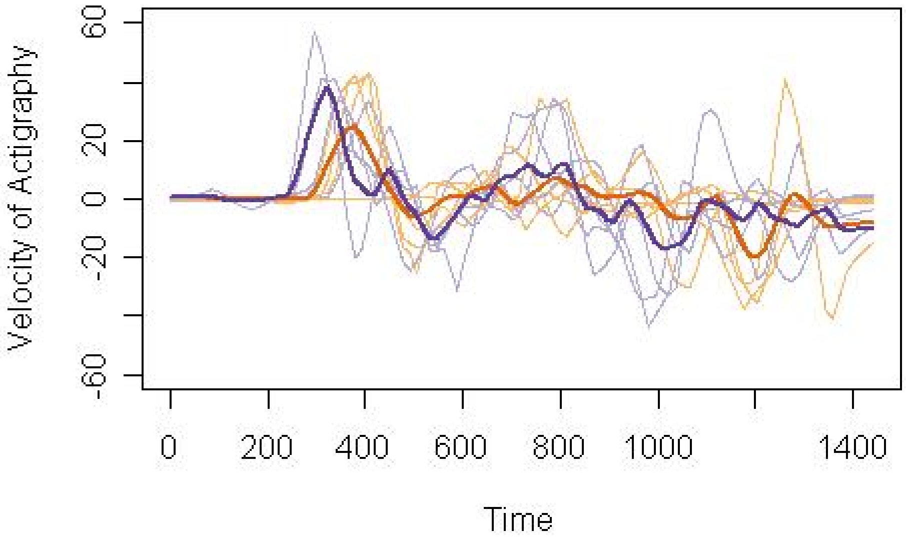
Raw Data

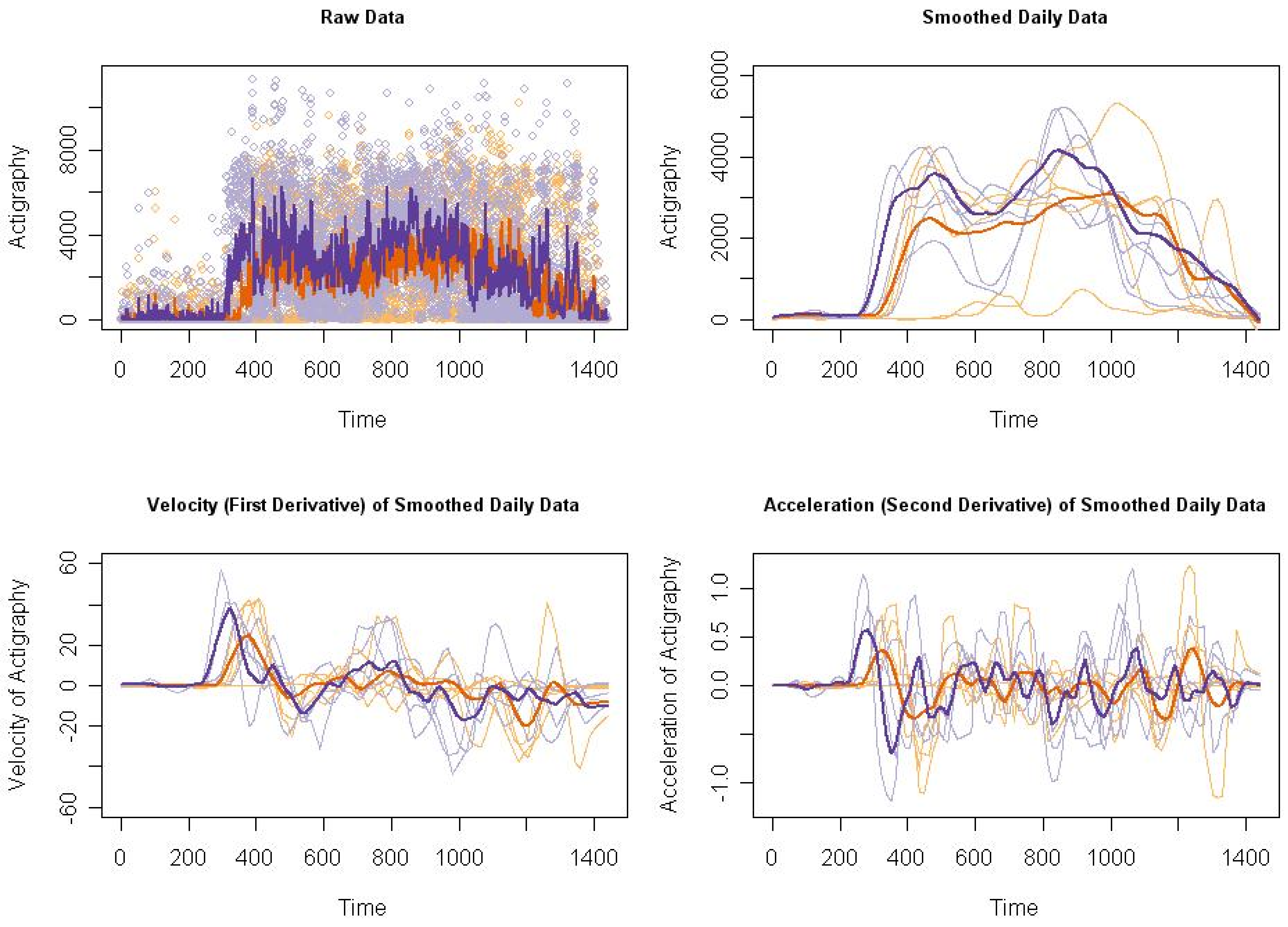


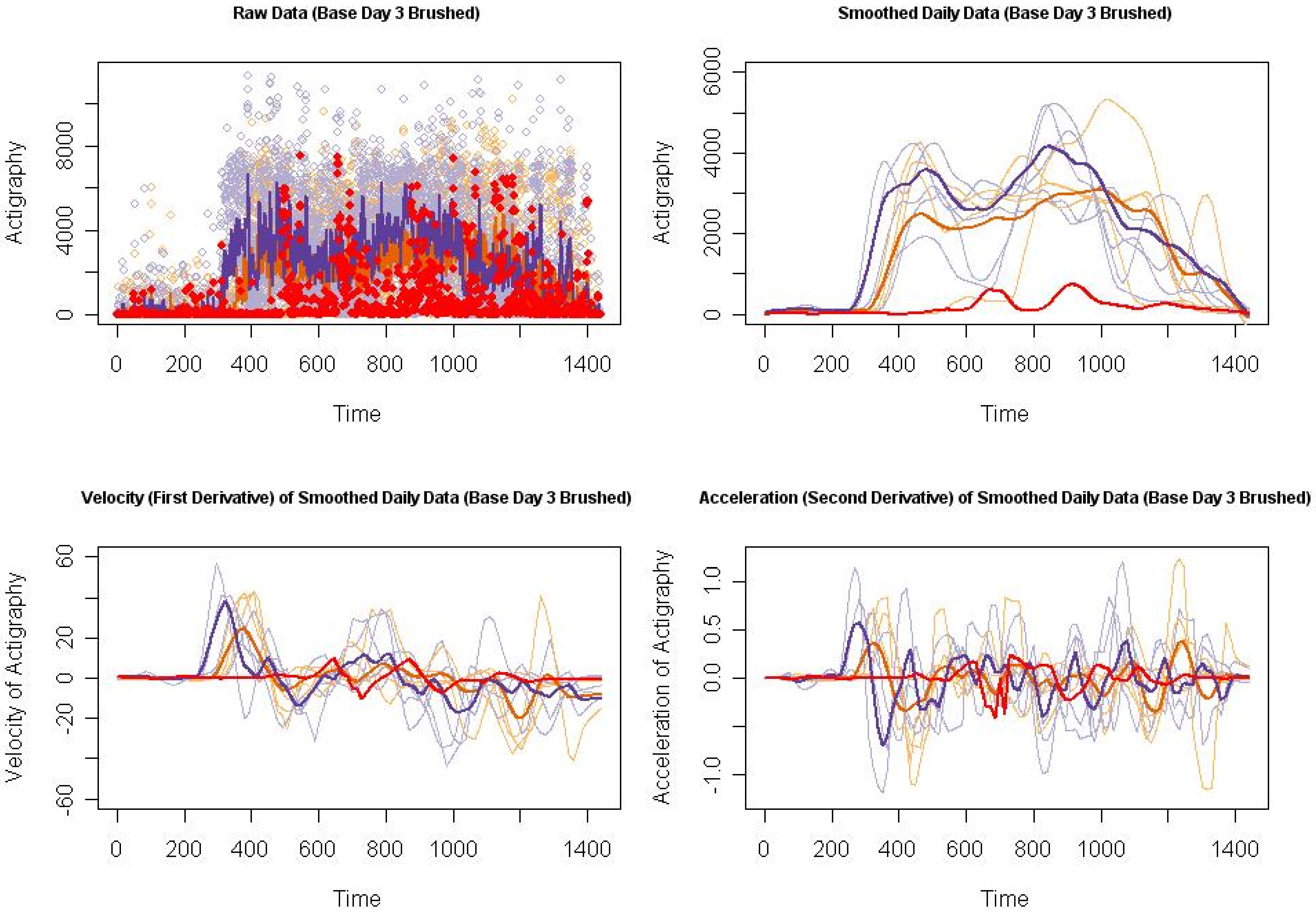
Smoothed Daily Data



Velocity (First Derivative) of Smoothed Daily Data




Raw Data
Actigraphy
0
4000
8000
0
200
400
600
800
1000
1400
Time
Smoothed Daily Data
Actigraphy
0
2000
4000
6000
0
200
400
600
800
1000
1400
Time
Velocity (First Derivative) of Smoothed Daily Data
Velocity of Actigraphy
-60
-20
0
20
60
0
200
400
600
800
1000
1400
Time
Acceleration (Second Derivative) of Smoothed Daily Data
Acceleration of Actigraphy
-1.0
0.0
0.5
1.0
0
200
400
600
800
1000
1400
Time

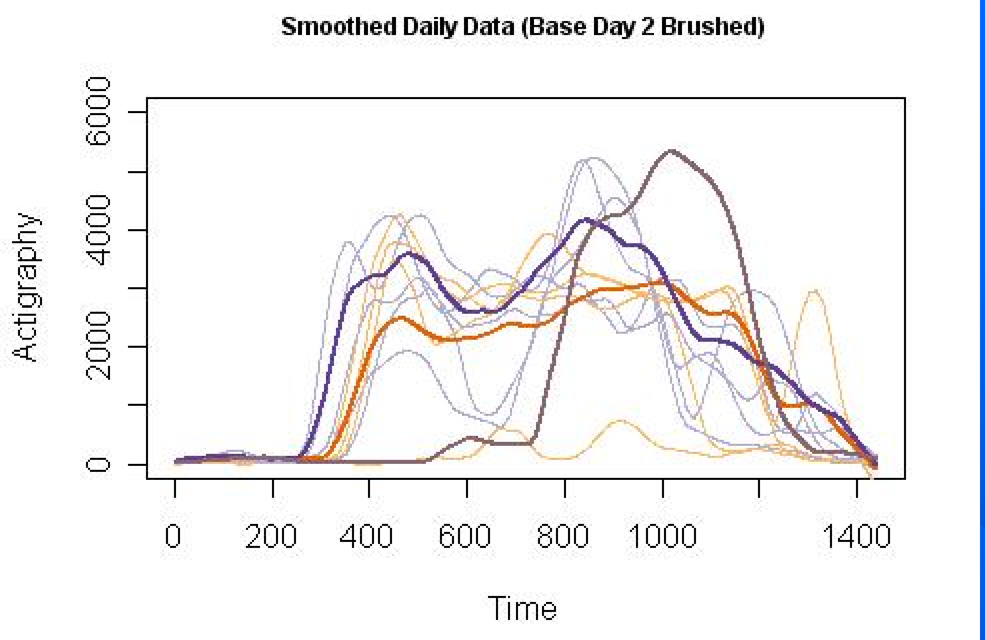


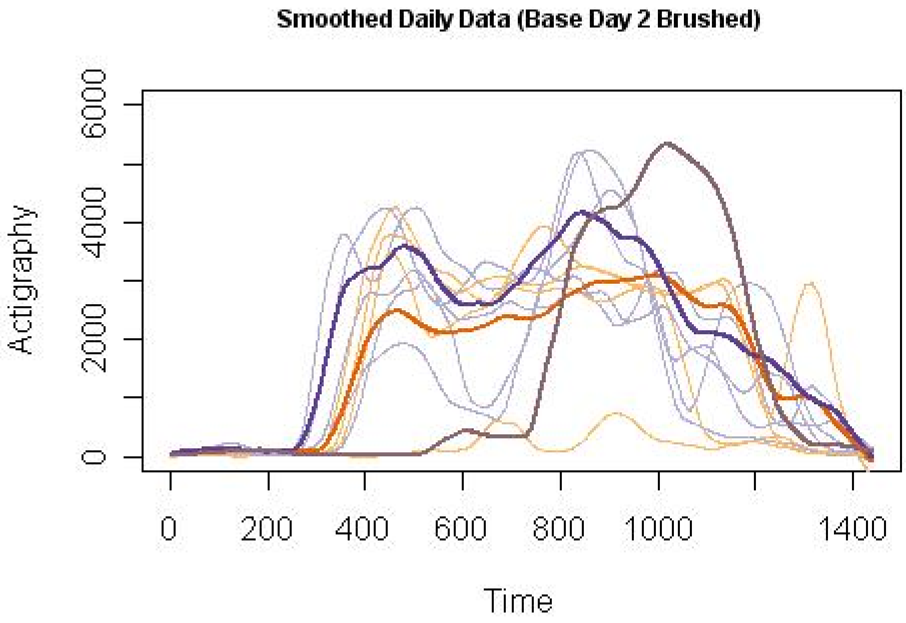
Raw Data (Base Day 3 Brushed)

Smoothed Daily Data (Base Day 3 Brushed)

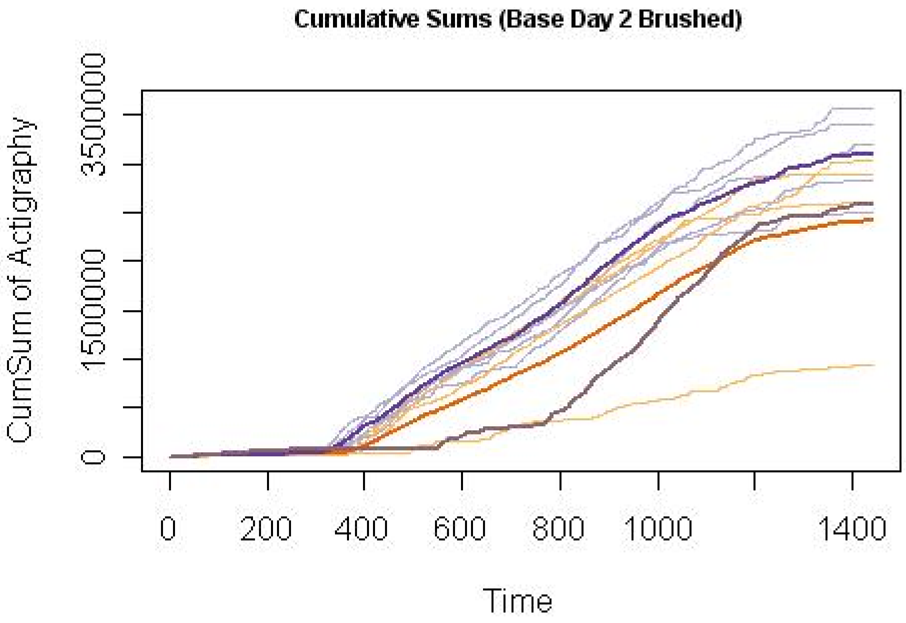
Velocity (First Derivative) of Smoothed Daily Data (Base Day 3 Brushed)

Acceleration (Second Derivative) of Smoothed Daily Data (Base Day 3 Brushed)

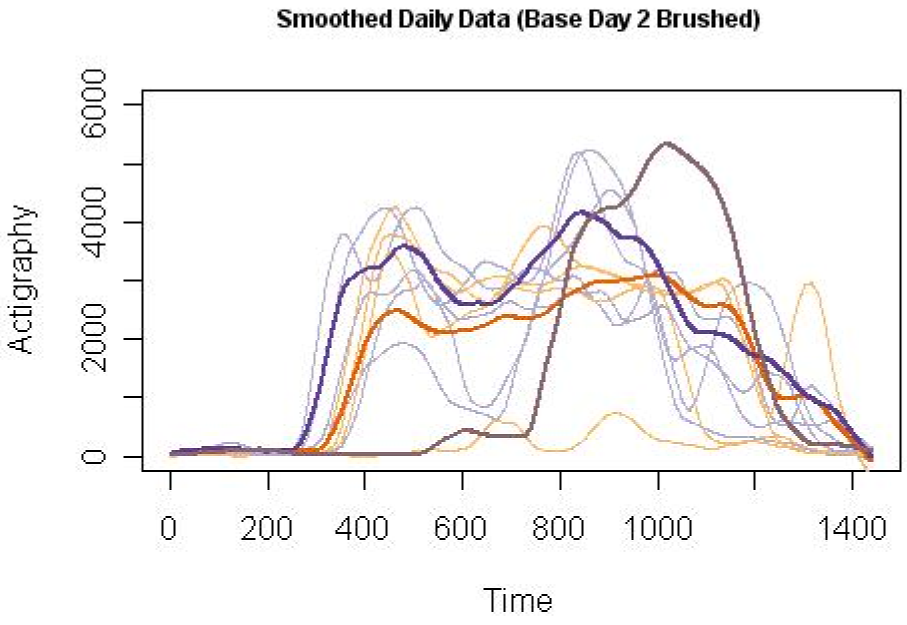

Smoothed Daily Data (Base Day 2 Brushed)
Actigraphy
0
2000
4000
6000
0
200
400
600
800
1000
1400
Time



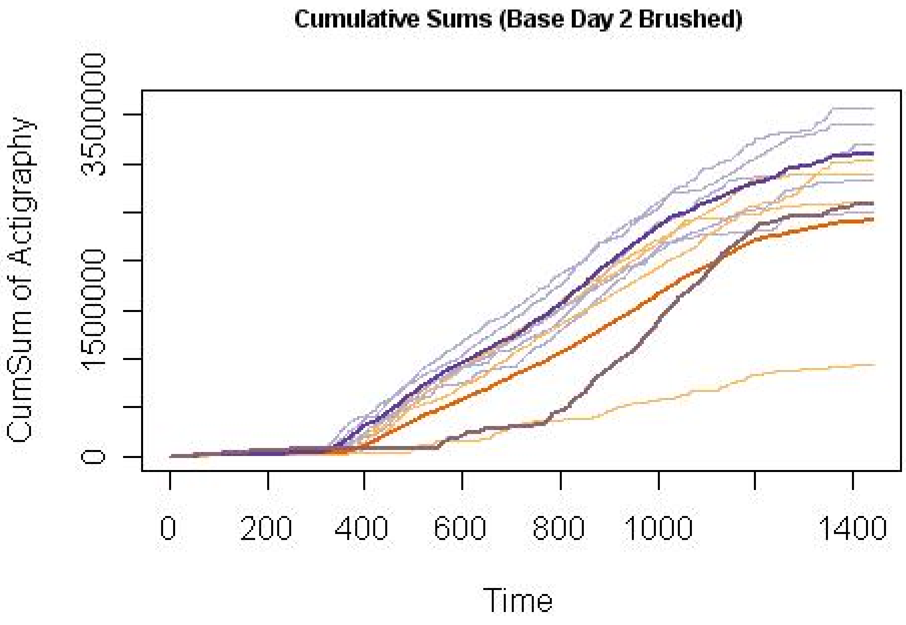
Smoothed Daily Data (Base Day 2 Brushed)



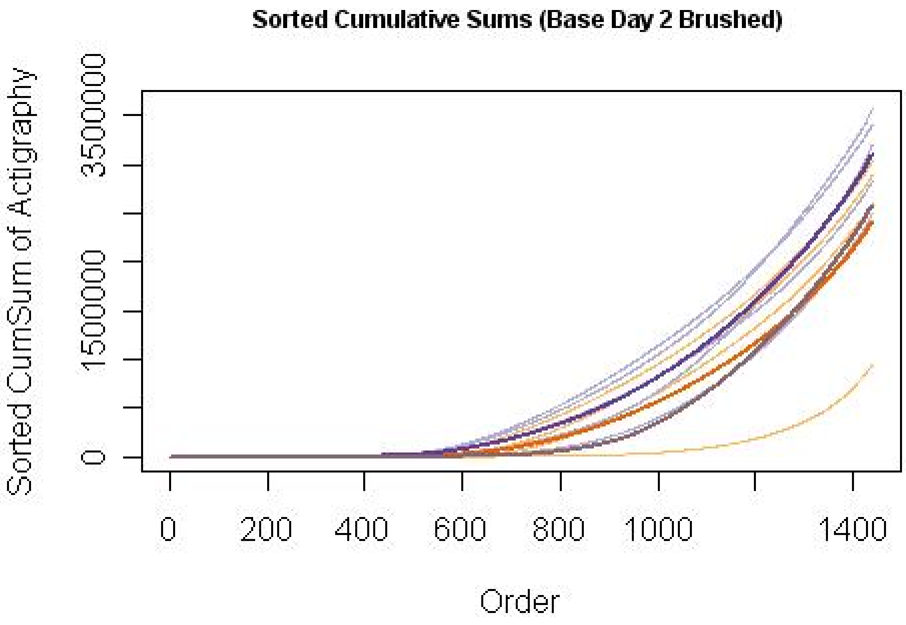
Cumulative Sums (Base Day 2 Brushed)



Smoothed Daily Data (Base Day 2 Brushed)



Cumulative Sums (Base Day 2 Brushed)



Sorted Cumulative Sums (Base Day 2 Brushed)

Conclusions

Visualization of Actigraphy Data provides

- Potential for application in various medical fields
- Additional insights into actigraphy data
- Ease to compare baseline and past-treatment data
  - » of a single patient
  - » of multiple patients
  - » to identify outliers
  - » to compare averages

Overall Conclusions

- Visual approach effective to see unexpected structure in data
- Combination of different techniques most effective
- Can be used for almost all types of data:
  - Educational Data
  - Social Sciences Data
  - Environmental Data
  - Medical Data
  - Economic Data (not shown here)

Questions ???

– or –

send e-mail to:

symanzik@math.usu.edu