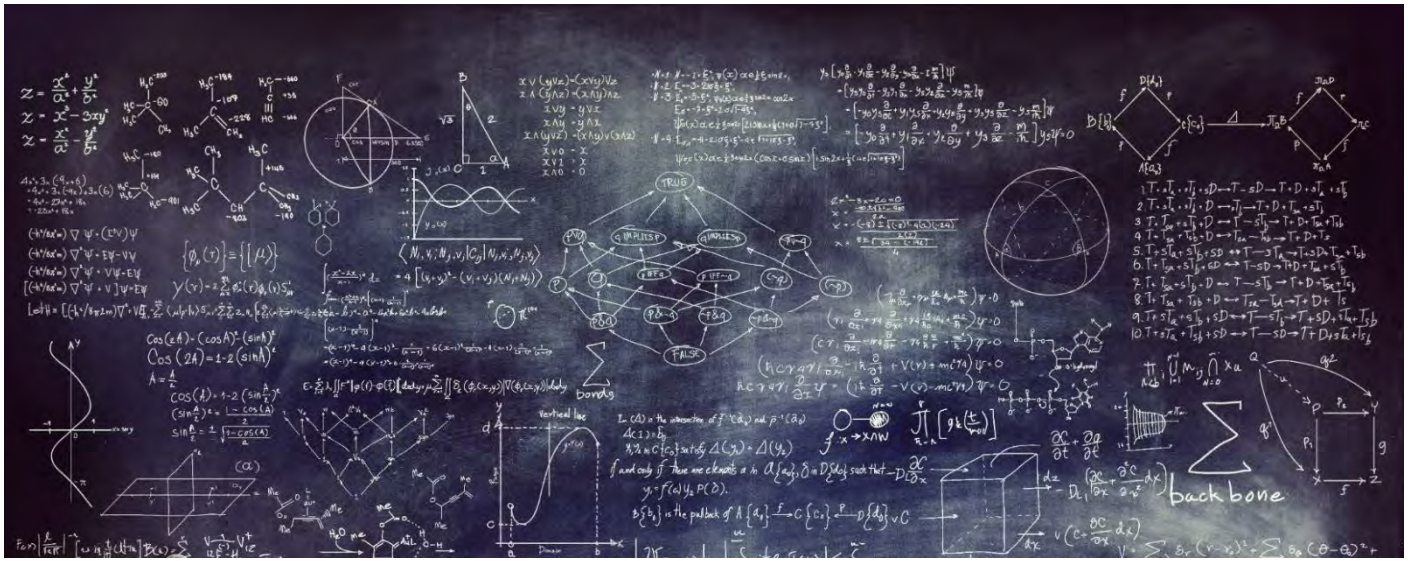


STEM

STEMS

STEAM

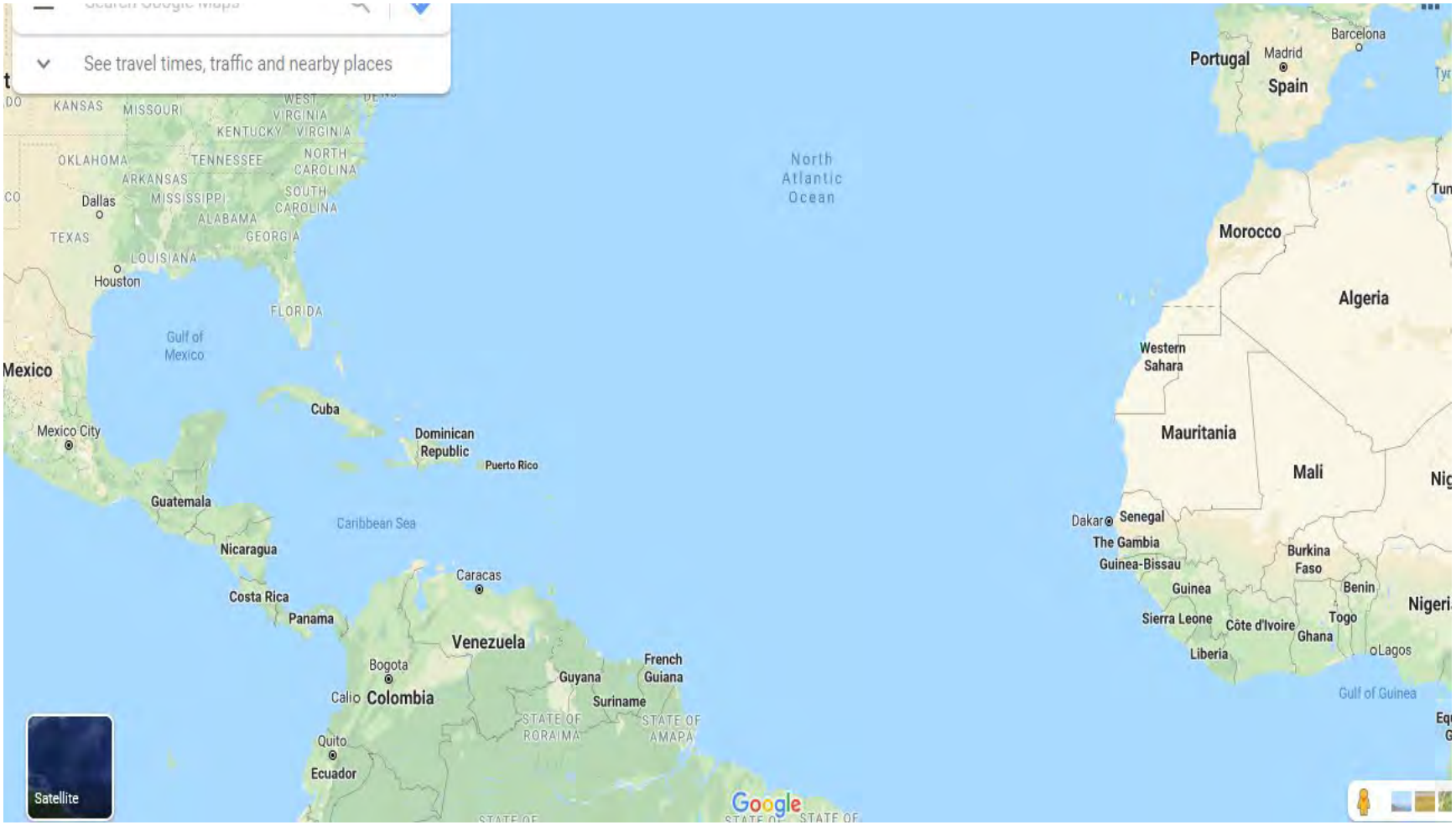
STEM H

STEM Topics in Data Science with Applications

Norou Diawara
Mathematics and Statistics Department, ODU

REYES, July 15th , 2020

STEM is worldwide



- Studied at University Cheick Anta Diop
- Then Universite du Havre, in France
- Then Auburn University
- Offered teaching position @ Old Dominion University

What is needed to succeed in a STEM field?

CRITICAL THINKING

PROFICIENCY IN SOLVING NON-ROUTINE PROBLEMS

TEAMWORK

Why is STEM needed?

SHORTAGE OF NURSES, DOCTORS, ENGINEERS, MATHEMATICIANS

STATISTICIANS

DATA SCIENTISTS

STEM JOBS WILL INCREASE

BETTER PAY THAN EVER

- Future
- Environmental
- Health
- Economics
- Behavior
- Accountability
- International Community

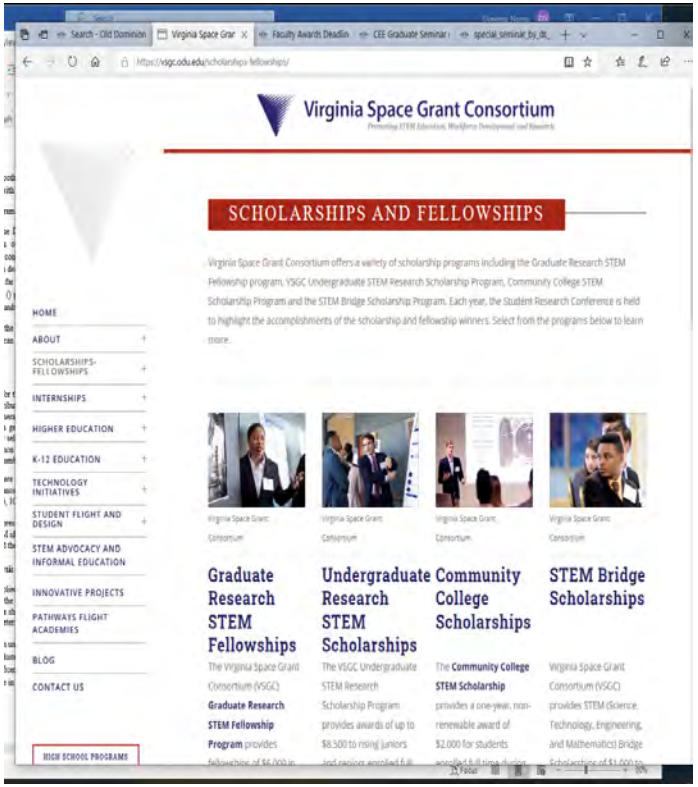
ROBOTS WILL SOON TAKE OVER MANY AUTONOMOUS JOBS

IMPROPER USE OF TECH CAN LEAD TO DISASTER

Resources?

Grants for STEM

- Advancing Informal STEM Learning (AISL) National Science Foundation. ...
- Science Education Partnership Award (SEPA) National Institutes of Health. ...
- Bay Watershed Education and Training Program (B-WET) ...
- Innovative Technology Experiences for Students and Teachers (ITEST) ...
- Laboratory Equipment Donation Program (LEDP)
- Universities and Colleges

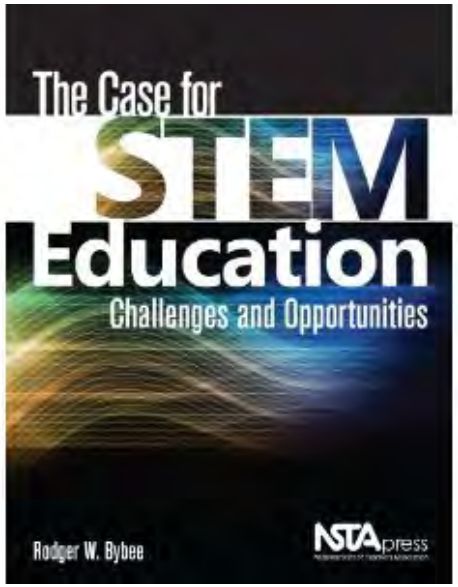


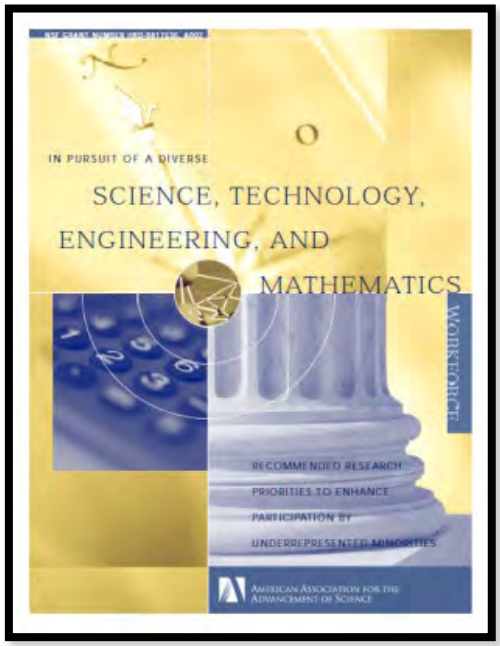
STEM

- Science
- Technology
- Engineering
- Math
- Statistics & Data Science → STEMS

STEM -sTEM- stEM- steaM

- Health → STEM-H
- Arts → STEAM





Numbers and data

Prime Numbers
Data Functions
Algorithms

Motivation

- Lifetime data two related events.

System with two components.

- Cannot assume independence.
- Positive support distributions.

Numbers and data

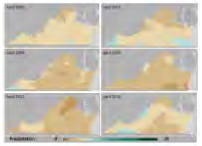
Data & design

ODU

R. A. Fisher: capture variations, causality
Blocking, Randomization, Replication, and Resampling
Resolution

$$Y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$$

- Moran, 1950
- Martin & Oeppen, 1975
- Stoyan & Stoyan, 1998
- Møller & Díaz-Avalos, 2010
- Valliant, et al., 2011



- Meddens & Hicke, 2014
- Neuroscience Division @ Ely Lilly
- Center for Biomedical Imaging Data @ Emory University

- Conny Palm
- Palm calculus

Interested in studying the properties of a chosen point of th

- conditional probabilities must be calculated
- the conditional distribution provides a flexible tool in patterns

Mathematician

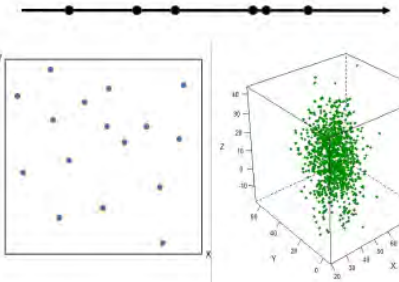
By ASSOCIATED PRESS · AUG 26, 2018

Share
Tweet
Email

Katherine Johnson

CREDIT NASA

A *point process* is a random point field with each point representing event of interest.



Problem of Interest

We consider two marginally gamma distributions X and Y. Put a linear relation thru a latent random variable (r.v.) Z independent of X.

More precisely, set

$$Y = aX + Z, \quad a > 0. \tag{1}$$

Applications

- Medical field

  pair of organs

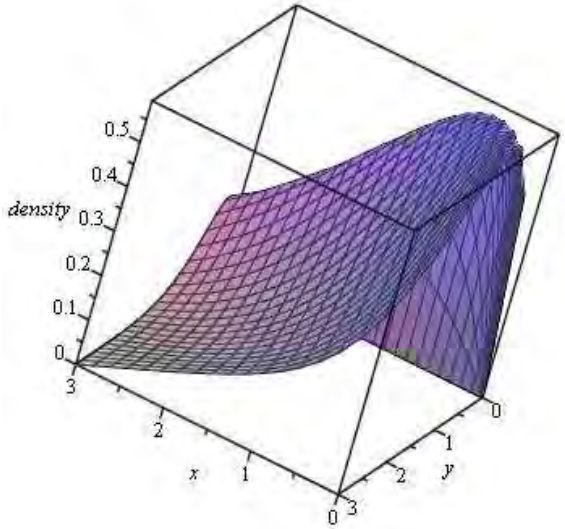
- Engineering

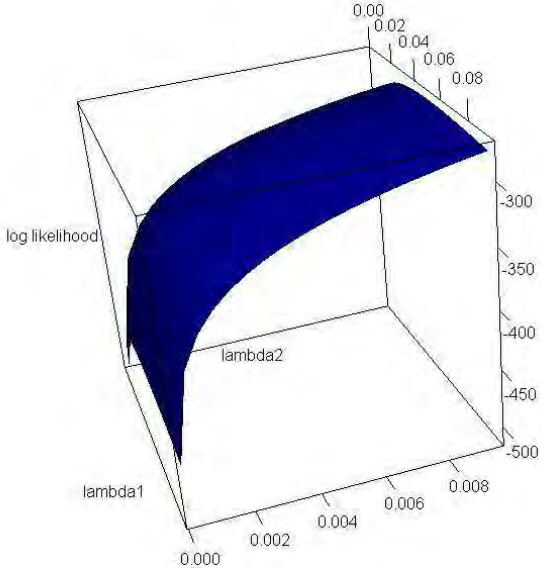
  failure of second component initiated by the first

Analysis

The distribution of Z takes different forms and contains information about the relationships between different variables' values.

1. Z is continuous gamma only when a takes a special form.
2. The expressions for Z
   - $\alpha_1 = \alpha_2 = \alpha \in \mathbb{N}$,
   - or $\mathbb{N} \ni \alpha_1 \leq \alpha_2$,
   - or $\mathbb{N} \not\ni \alpha_1 < \alpha_2$ with $[\alpha_1 + 1] = N \leq \alpha_2$.





Forecasting under Structural Changes in Non-linear Time Series

Motivation

- Partially observed time series models are studied under various condition,eg. State Space Models (Durbin and Koopman, 2001), Dynamic Models (West and Harrison, 1997), and Hidden Markov Models (Cappe et al. 2005).
- Challenging-Formulation of conditional distribution and reconstruction of the unobserved components from the observed values.
- More challenging- if the variance is not constant, changing over intervals.
- Common in financial, economic and environmental time series data.

Forecasting

- Forecasting for multiple regression where different covariates have different structures (Gagnon and Kedem, 2010); Forecasting based on Bayesian approach assuming possibility of breakpoints in the forecast horizon (Pesaran, Pettenuzzo and Timmermann , 2005)
- We developed the method based on joint density of mixture distribution of combined error data.
- Parameters are estimated by using EM algorithm.
- Estimation is improved by using bootstrap combined with EM algorithm.

Parameter Estimation Using EM Algorithm

- General case: Mixture of r density functions

$$Let\ \zeta \sim f(\varsigma|\mathbf{p}) = \sum_{j=1}^{r} p_j f_j(\varsigma)$$

$$logL(\mathbf{p};\varsigma) = \prod_{i=1}^{n} f(\varsigma_i|\mathbf{p}) = \sum_{i=1}^{n} log\left(\sum_{j=1}^{r} p_j f_j(\varsigma_i)\right) --> Not\ solvable\ in\ closed\ form$$

Define

$$W = (w_{ij}, i = 1,...,n; j = 1,...,r) = \begin{cases} 1, & \text{observation } \varsigma_i \text{ comes from the distribution } f_j \\ 0, & \text{else} \end{cases}$$

Log- Likelihood for complete data: ,

$$logL_c(\mathbf{p};\varsigma,w) = log\prod_{i=1}^{n}\prod_{j=1}^{r}(p_j f_j(\varsigma_i))^{w_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{r} w_{ij} logp_j + \sum_{i=1}^{n}\sum_{j=1}^{r} w_{ij} logf_j(\varsigma_i)$$

Conditional Expectation $E_{p_j^{(k)}}(w_{ij}|\varsigma) = P_{p_j^{(k)}}(w_{ij} = 1|\varsigma) = w_{ij}^{(k)} = \frac{p_j^{(k)} f_j(\varsigma_i)}{\sum^{r} p_j^{(k)} f_j(\varsigma_i)}$.

Spatio-temporal models

- Goal: model for random patterns of points in a high dimensional space (locations) where events of interest (may them be cyber-attacks, contamination sources, etc...) are observed, are accounted for under such processes, with addition of time frequencies.
- Which measure? Moran's statistics

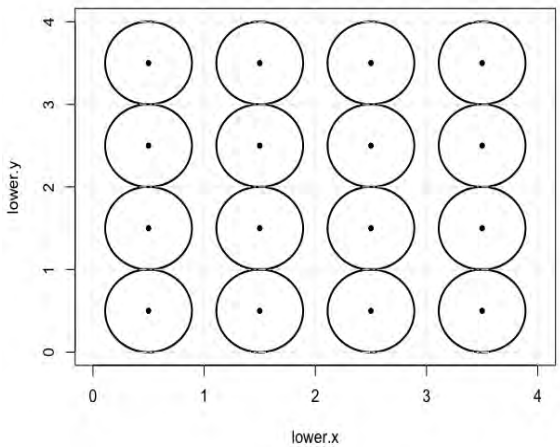
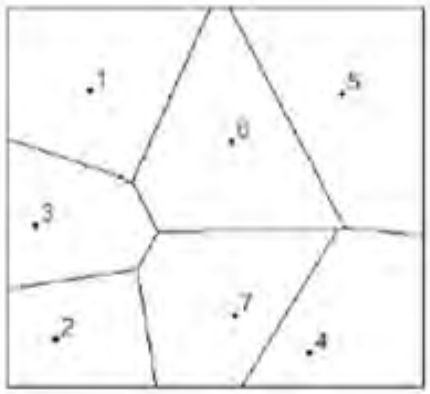


Figure 1: Regular space with 16 subareas



(a)
Figure 2: Poisson point process

Spread under regular spacing

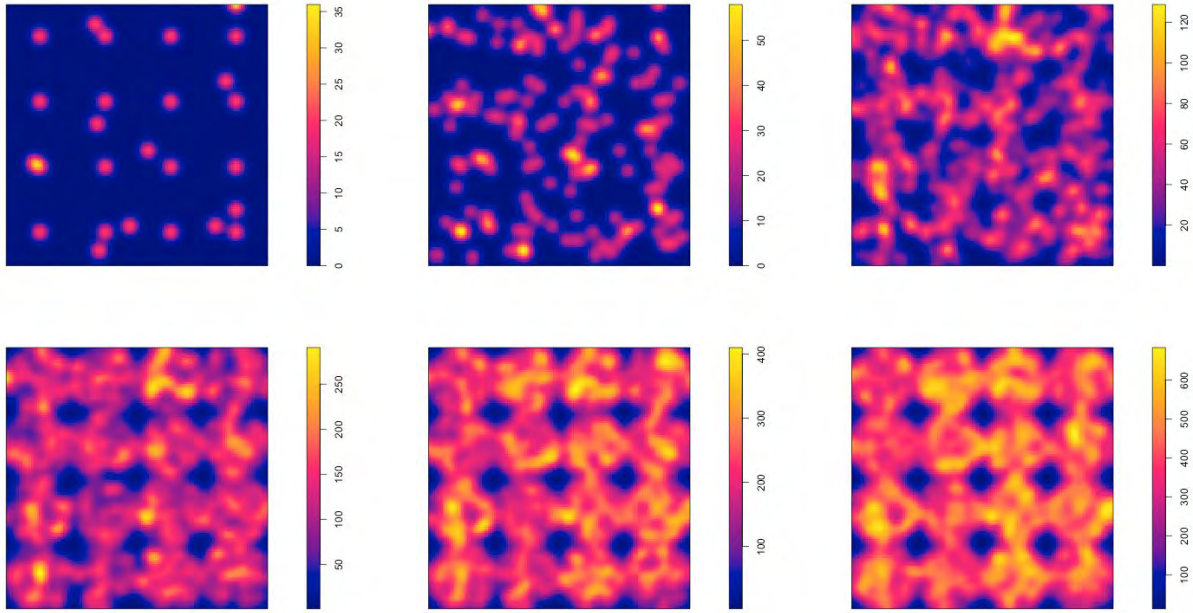


Figure 3: Apparent spatial clustering of regular space (16 subareas) through t = 6

Evolution of system

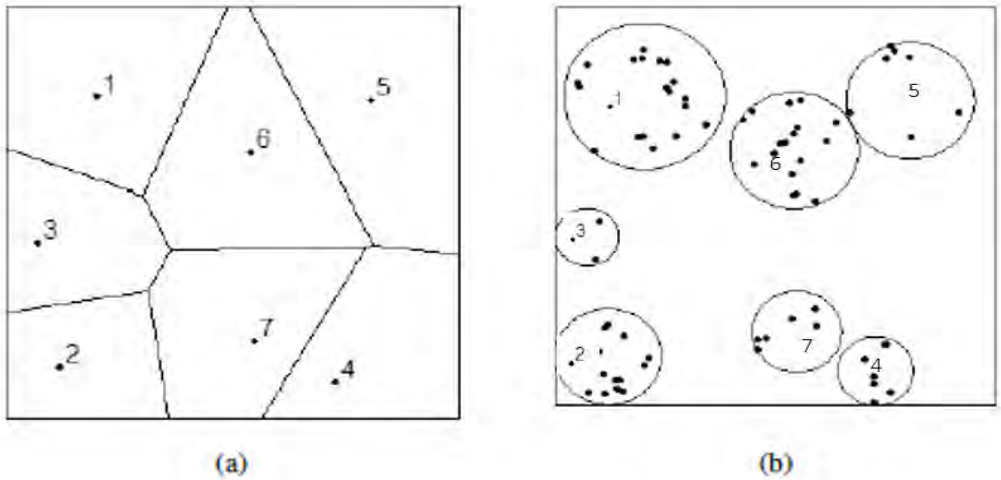


Figure 4a: Poisson point process plots at (a) t = 1 and (b) t = 2

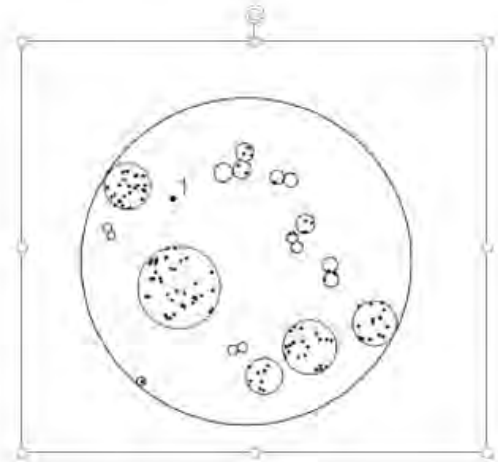


Figure 4b: Poisson point process disc 1 at t = 3

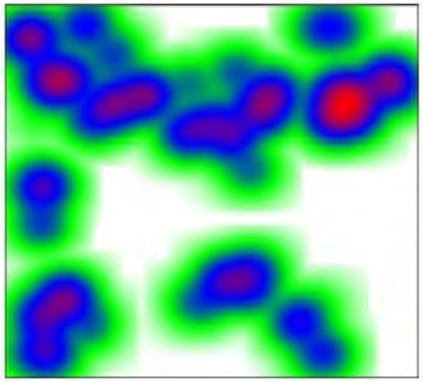
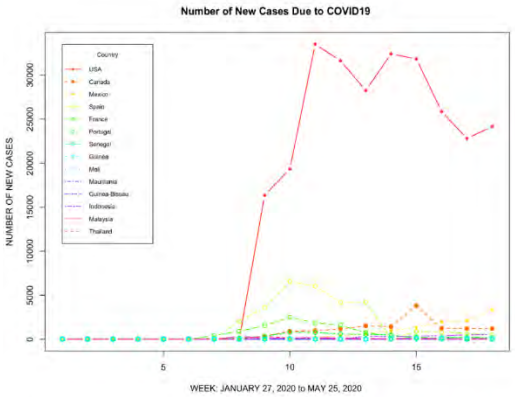


Figure 4c: Apparent spatial clustering through t = 4

COVID-19 new cases



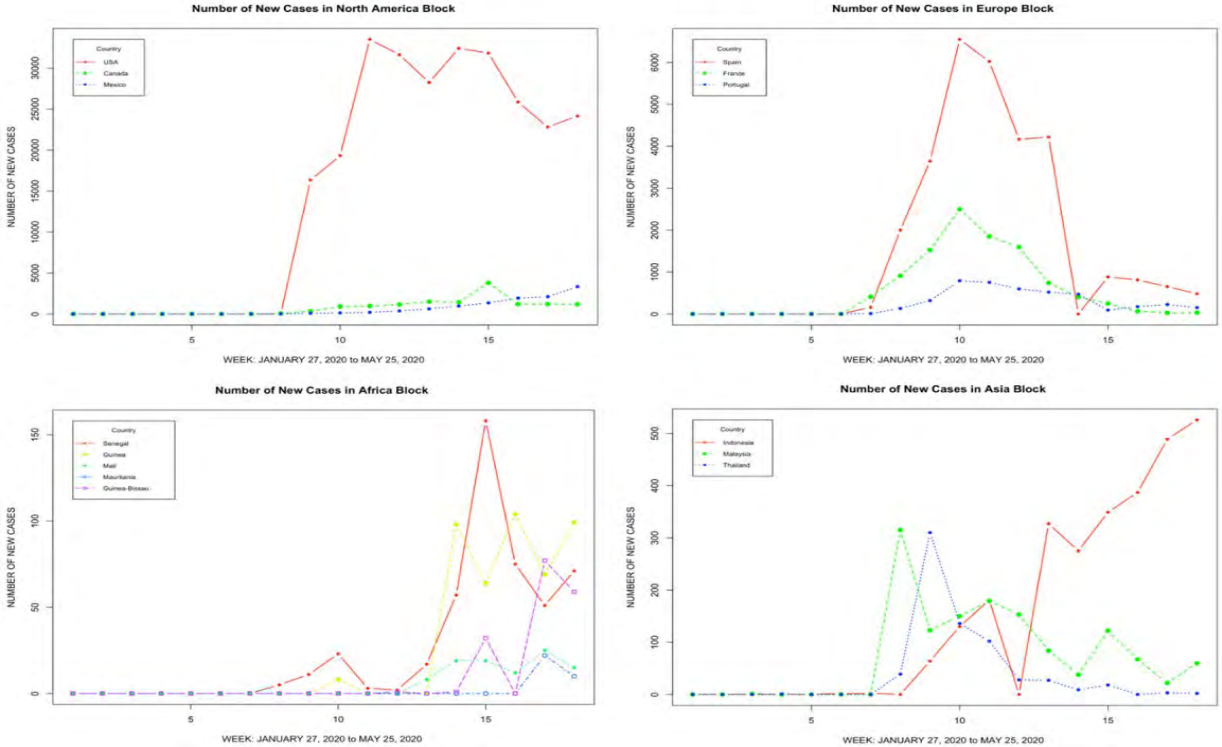


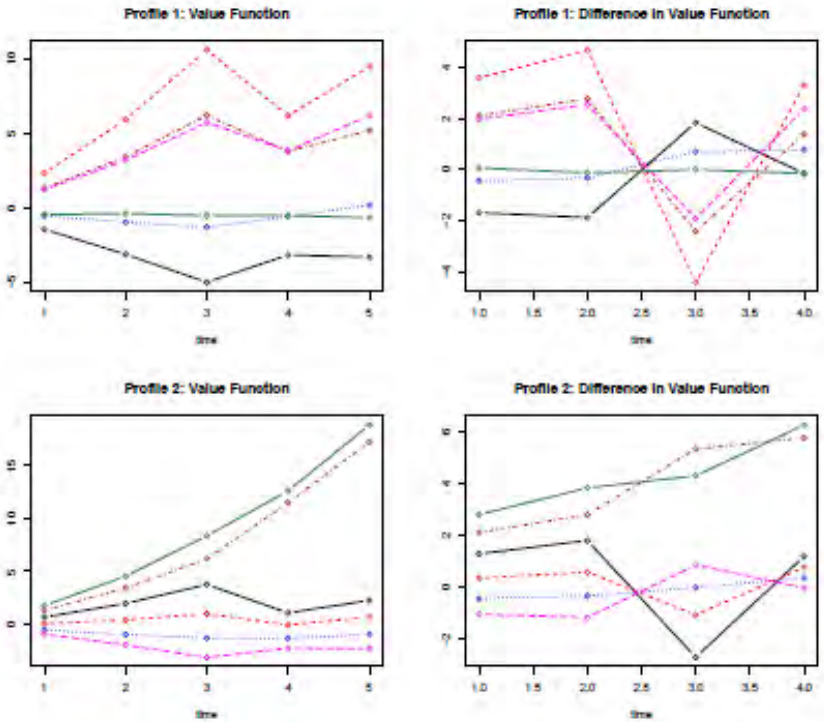
Figure 5: Line plots of new cases per block

COVID-19 new cases

Figure 6: Shapes / Colors of new cases per block

Time Dependent Consumer Behavior:
Attribute and Attribute-level Best-worst DCEs

- The use of Markov Decision Processes (MDPs) allows us to project the future discounted utilities for the different attribute-level best-worst pairs. Data is simulated from a real-world example and the discounted utilities for the different attribute-level pairs are projected across five time periods.
- Patterns in the discounted utilities suggest that some choice pairs will continuously yield high returns while others yield negative returns. We find that the model of individual choice preferences yields to a balance from several factors/controls at hand.



Other Applications

- Driving safety: the seat belt data
- USAID data (finance)
- Brain functions
- Calculating patterns of certain pandemics
- And more

Seat Belt Use in Virginia

Predicting Seatbelt Use under a Higher Order Structure Design in 2017-2019 Survey Design

Weighted Item Response Theory Model

Figure 7a. Comparisons of seat belt use by year for males and females.

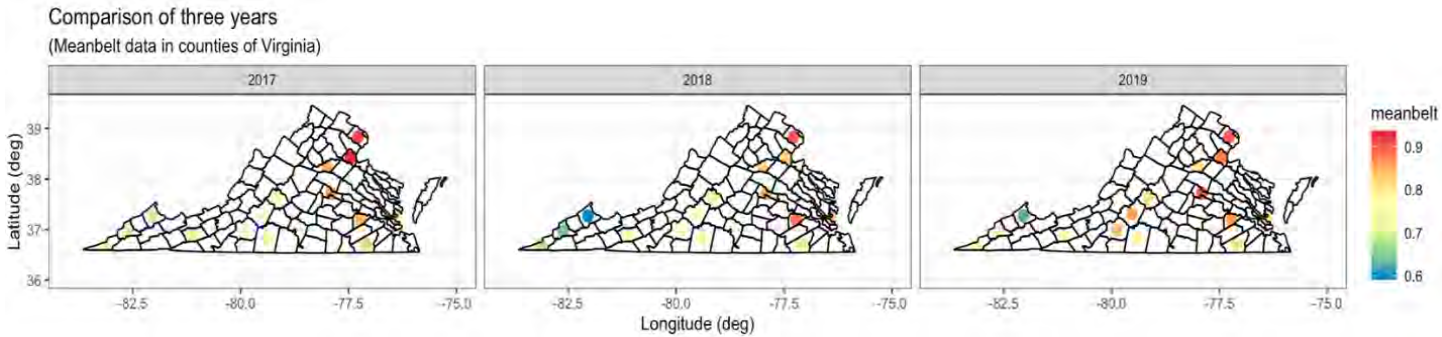


Figure 7b. Comparisons of seat belt use by year for males.

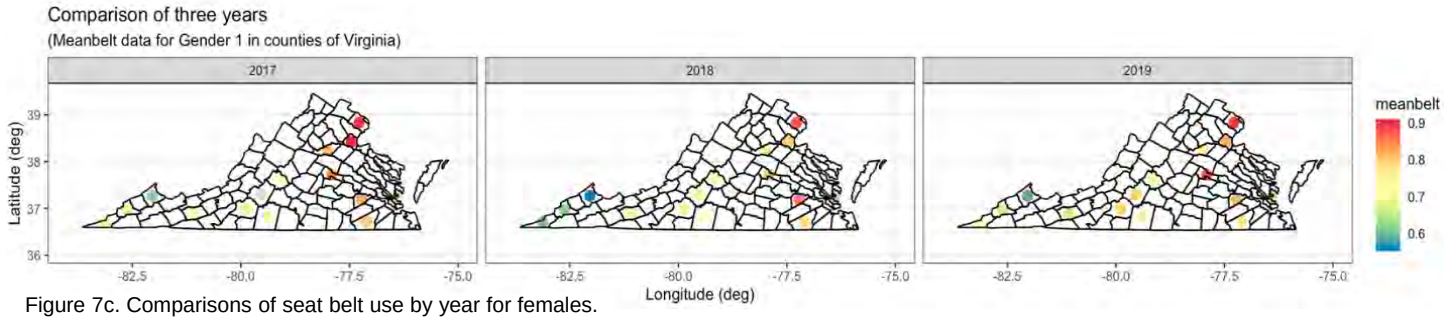
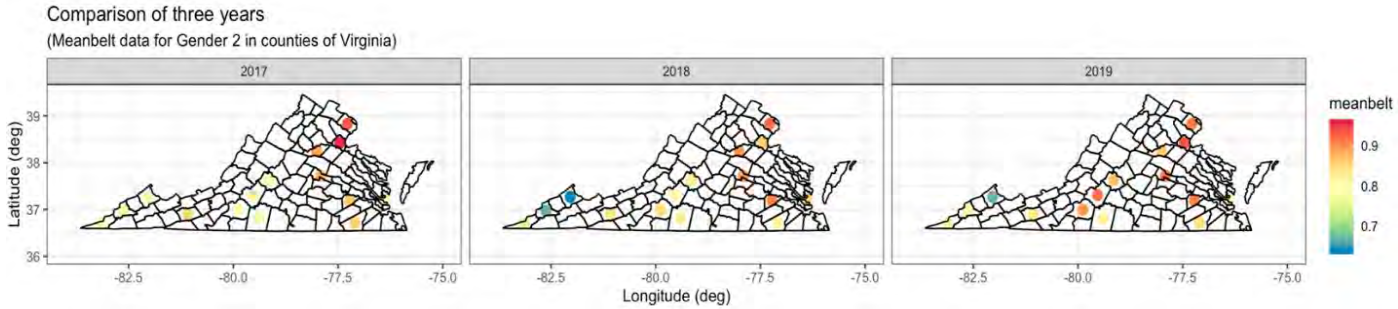


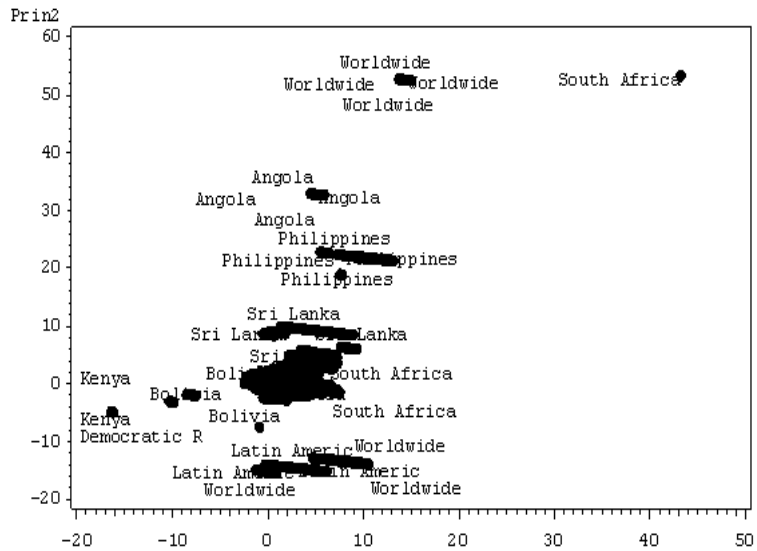
Figure 7c. Comparisons of seat belt use by year for females.



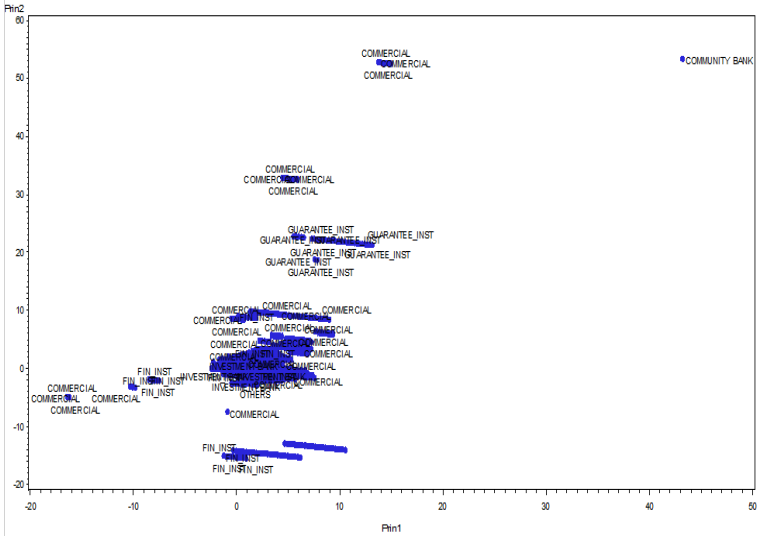
REVIEW OF USAID/DCA DEFAULT HISTORY AND DEVELOPMENT OF DEFAULT ASSUMPTIONS

The data for this project has been provided by USAID's Development Credit Authority (DCA). It is a panel data set from more than 73 countries and consists of 56 variables in 548,309 data rows, from 1992 to 2012.

$$Default\ rate = \frac{Adjustment\ Factor * Claim\ amount}{Transaction\ Amount}.$$



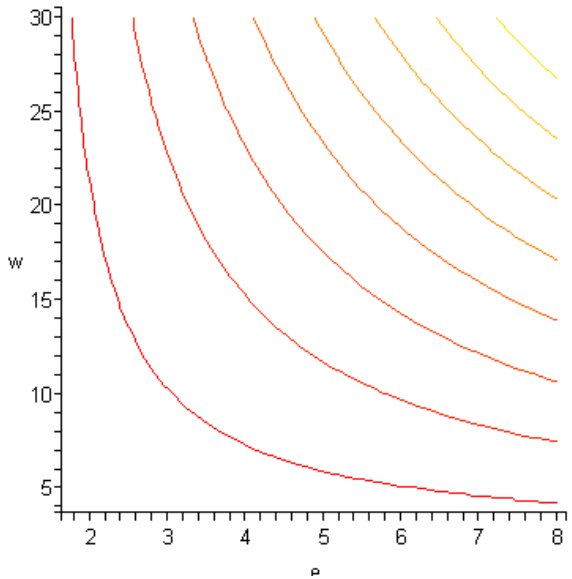
Plot of the two primary principal components by country



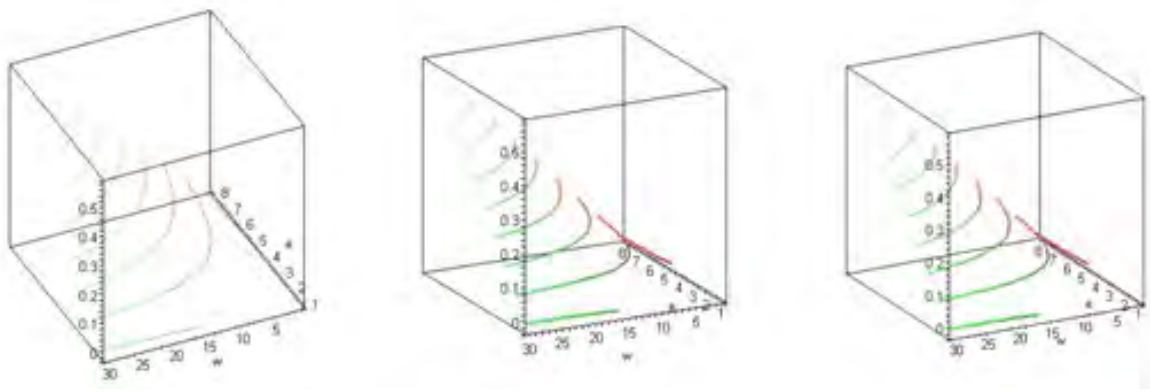
Plot of the two principal components by lender types

WARF Score, Event Period, Loan Duration and Origination Fee Percent, and then

$$r = \beta_0 + \beta_e e + \beta_w w + \beta_l l + \beta_f f + \varepsilon$$



Profile plot of default rate with respect to Event Period (e) and WARF Score (w)



Three-dimensional plot of the default rate based on WARF Score (w) and Event Period (e)

Maps of the region showing the key pairwise differences between countries.

The maps were produced in QGIS, an open source GIS software equivalent to ArcGIS. Colors were added to the m using Microsoft Paint.

Unmodified map color in light pink – no data available for those countries. Yellow countries have USAID data.

Original map colors represent countries without USAID data. White represents countries that had no defaults in their records (not included in the analyses). Yellow represents countries that had at least 1 defaulting event across records.

Figure 3: Original and white colors represent the same classifications Here, yellow represents countries that had less than 1000 observations and at least 1 defaulting event – no significant differences were found between these 4 countries. Green represents countries that had more than 1000 observations and at least 1 defaulting event. The colored boxes around the green countries represent the absence of pairwise differences in least-squares means of countries as per the Bonferroni adjustment.

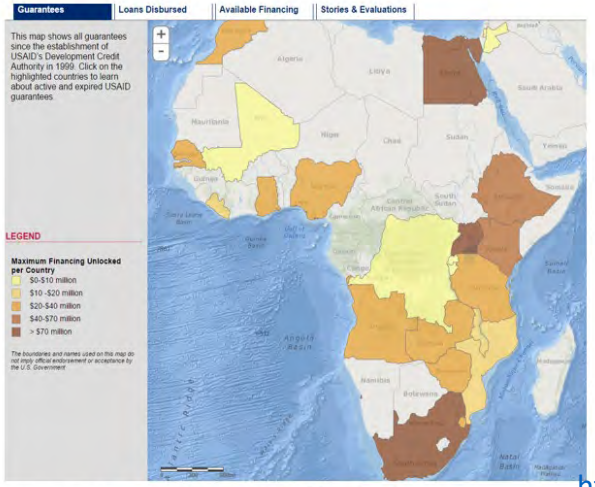


Figure 4: USAID map for all guarantees as per http://storymaps.esri.com/stories/usaidcredit/

https://stories.usaid.gov/usaidmap/?location=africa

Brain image and functions

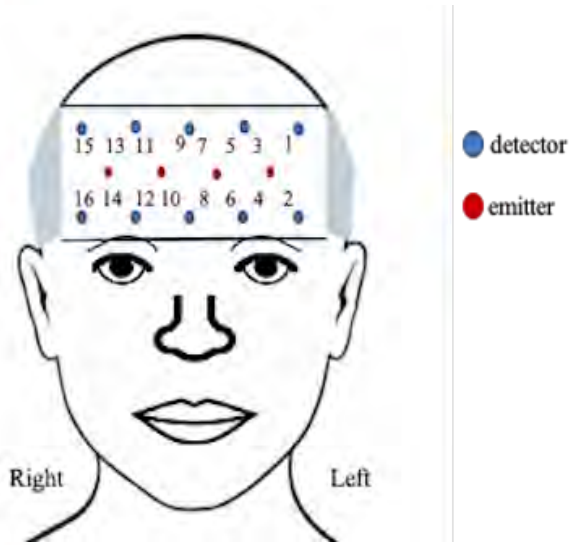
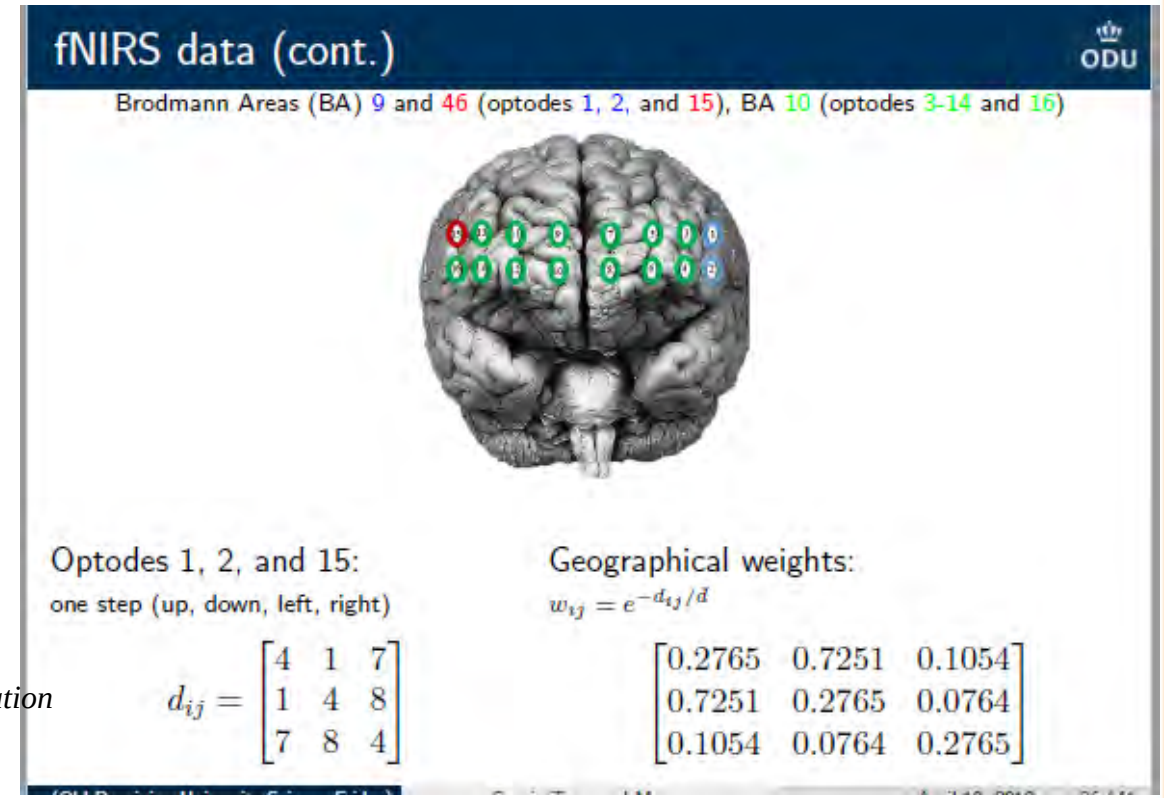


Figure 9a: Functional near-infrared configuration



Brain image and functions

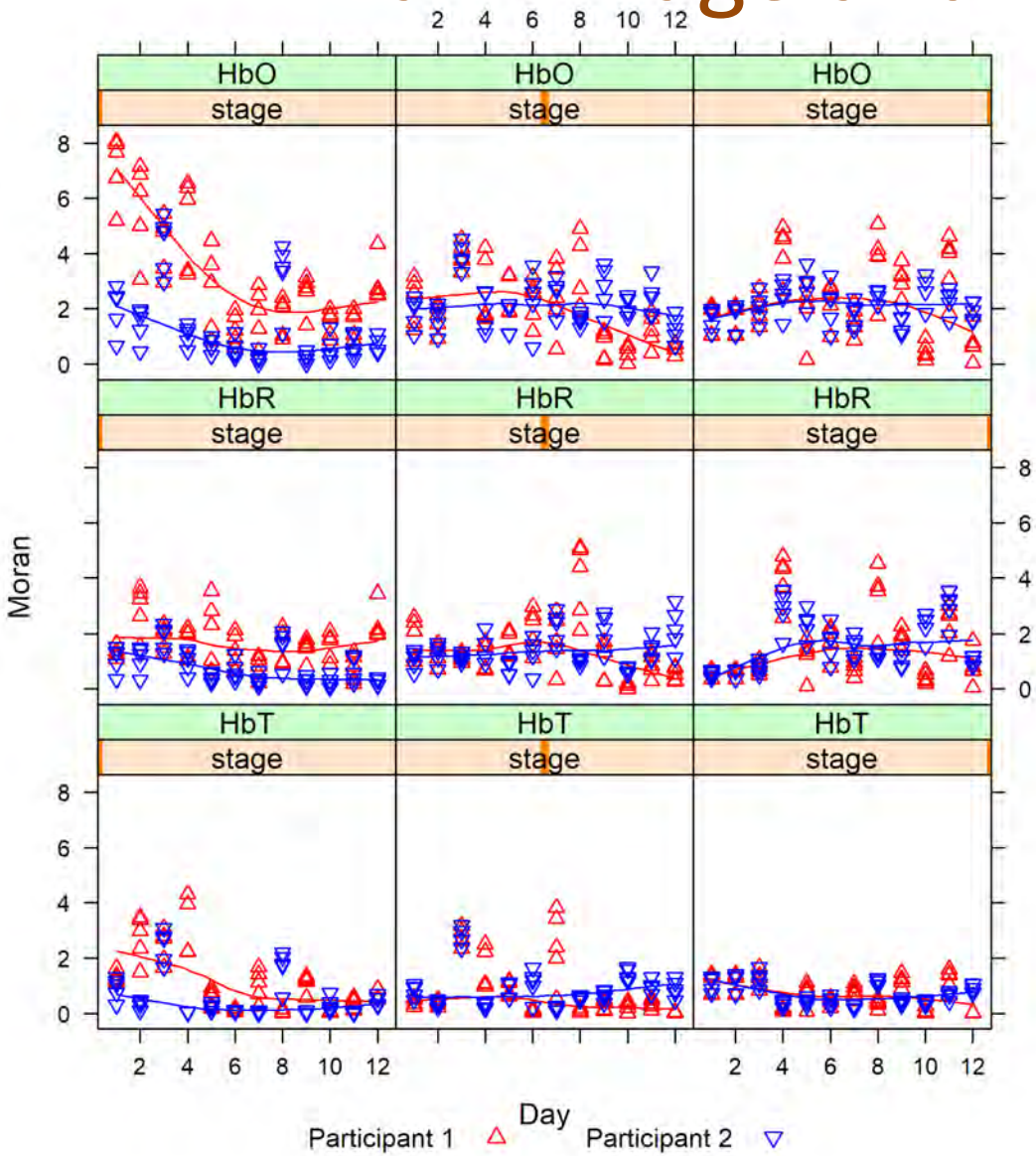


Figure 9b. Moran statistic for Grouping 1 for HbO, HbR, and HbT by stage for days 1 to 12.

Conclusion

Science really rocks

We need each of you

Thank you

Thanks to Dr. González
Thanks to REYES organizers

Support STEM

Questions?

Comments?

Thoughts?

Useful links

- Maple and Mathematica:
  - www.maplesoft.com
  - https://www.wolfram.com/mathematica/?source=footer (and https://www.wolfram.com/)

- R and SAS Software (free options)
  - https://www.sas.com/
  - http://www.r-project.org/
  - https://www.rstudio.com/products/rstudio/download/
-
- Data science challenge: Kaggle
  - www.kaggle.com