Application of "dominance analysis" with health data: Identifying regional spreaders of a viral pathogen and their socio-spatial predictors in Nova Scotia

Yigit Aydede (Saint Mary's University)
Jan Ditzen (University of Bolzano)

1st Applied Microeconomics Workshop
MLPortal - CLARI
August 2022

Motivation

- The spread of viral pathogens is inherently a spatial process.
- While the temporal aspects of a viral spread at the epidemiological level have been increasingly well characterized for a region, the spatial aspects of viral spread are still understudied.
- We can "reliably" identify "hotspots" by Spatial Scan Statistic (SSS - Kulldorff, 1999), but there is no study to investigate the directional spatial network of a viral spread: **what are the "spreader" regions, which affect a large number of other regions and connected to many others in a regional network?**
- Characterizing these spatial dynamics and understanding the factors driving them are important for anticipating local timing of disease incidence and for guiding more informed control strategies.

What's missing

- While a number of studies have examined individual-level risk factors for COVID-19, for example, few studies have examined geographic **hotspots** and community drivers associated with spatial patterns in local transmission.
- For example, while much research focus on the epidemiological and virological aspects of a transmission, there remains an important gap in knowledge regarding the drivers of geographical diffusion between places.
- Most existing studies use disease clusters (by SSS) which do not reveal the network of spatiotemporal spread of a viral pathogen.

Spatial Scan Statistic

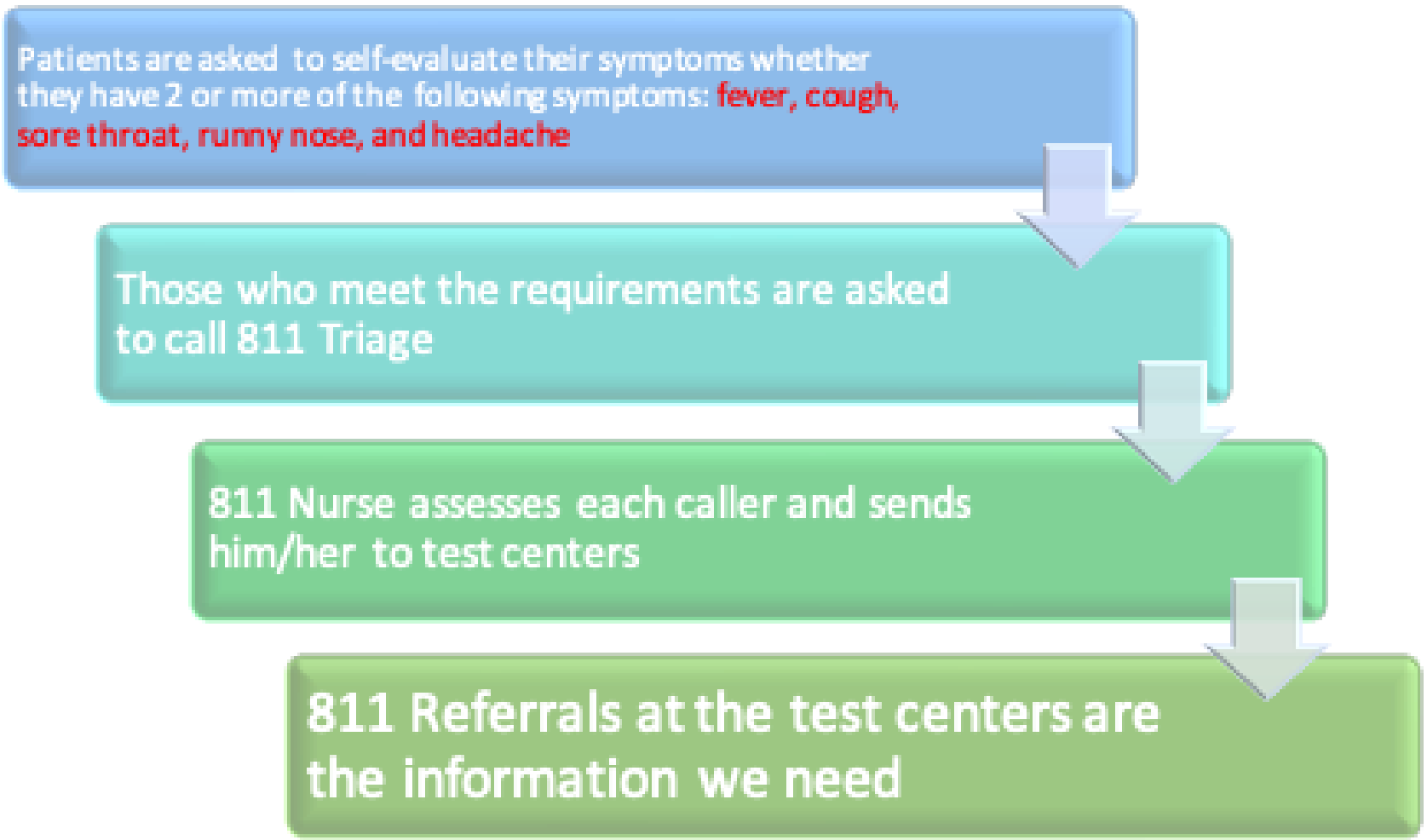
- A common problem in spatial statistics is whether a set of points are randomly distributed or if they show signs of clusters or clustering.
- Hence, a cross-sectional dispersion of disease may not be informative in terms of the disease clusters: a low-indecent region could be a cluster (hotspot) or a high-incident region could not be characterized as a hotspot (even if we use densities)
- The spatial scan statistic (Kulldorff, 1999) commonly used to detect "non-random" spatial disease clusters in epidemiological studies.
- These metrics evaluate whether a disease is randomly distributed or tends to occur as clusters over space based on cross-sectional information.
- PS: Toshiro Tango (2021): Spatial scan statistics can be dangerous.

Objective

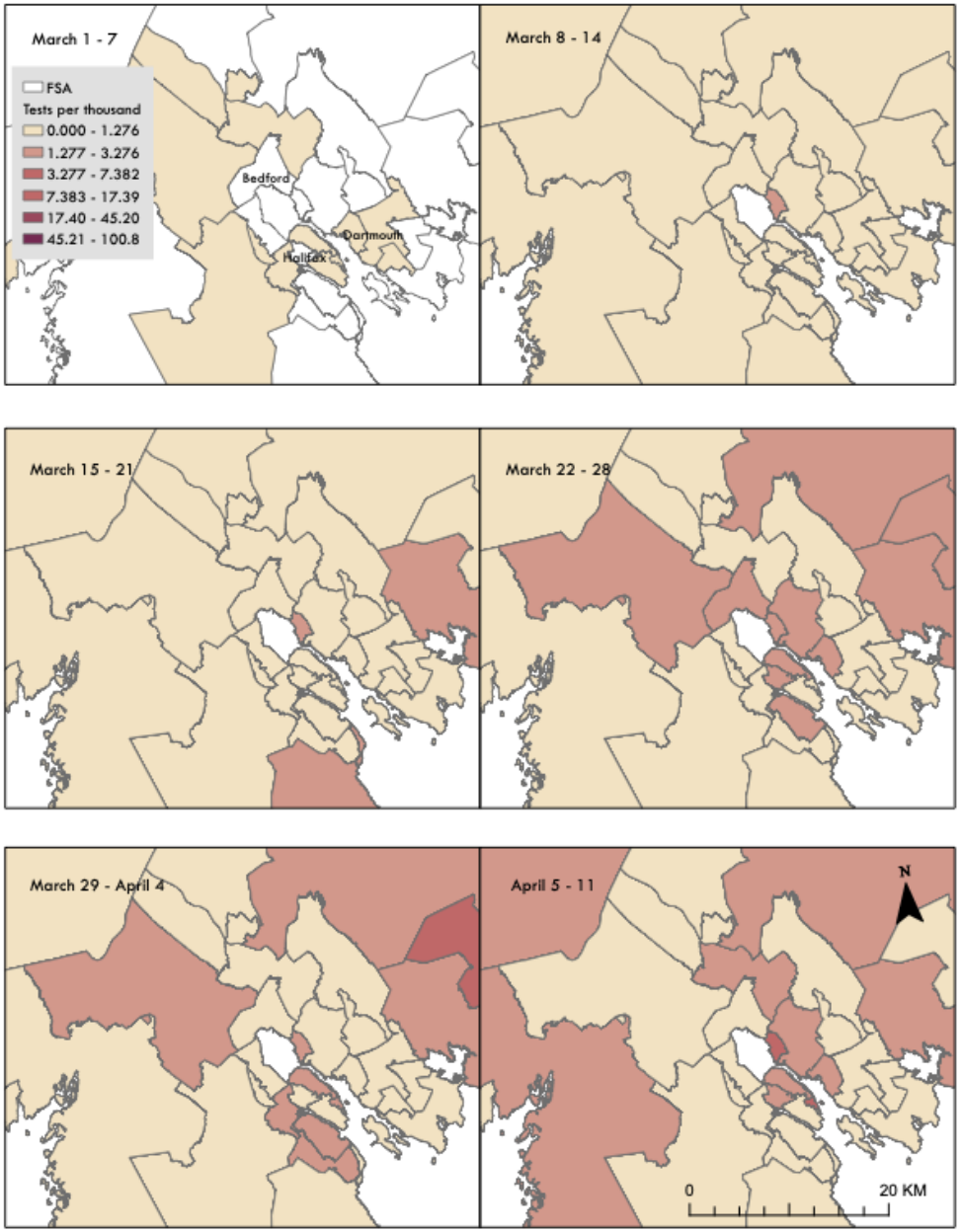
1. With spatio-temporal dynamics, to recover the network of a viral spread where the regions in their dominance are identified and ranked.
2. To find the socio-spatial predictors of being a “spreader” region.

Data on the viral spread

Thanks to a strict triage procedure in the first 5 months, we are able to use test numbers of COVID-19 reflecting a local viral spread (of other types of betacoronaviruses) and its spatial variation



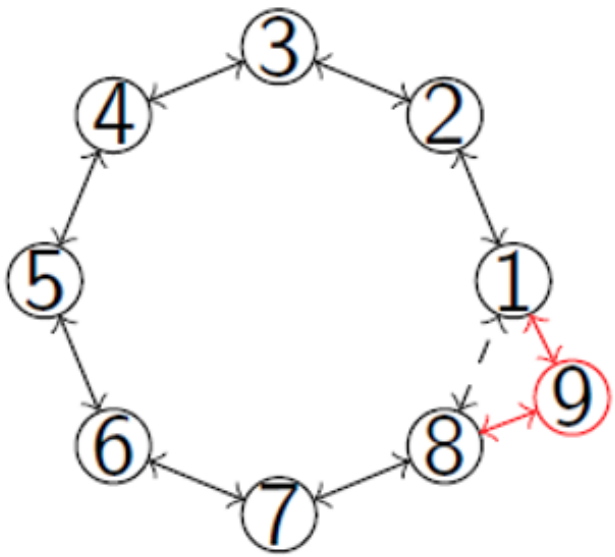
Heatmaps

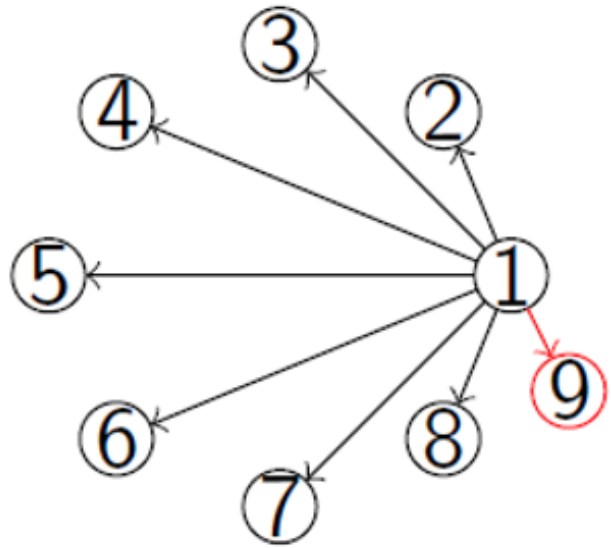

March 1 - 7
FSA
Tests per thousand
0.000 - 1.276
1.277 - 3.276
3.277 - 7.382
7.383 - 17.39
17.40 - 45.20
45.21 - 100.8
Bedford
Dartmouth
Halifax
March 8 - 14
March 15 - 21
March 22 - 28
March 29 - April 4
April 5 - 11
N
0
20 KM

Dominant units

- Dominant units are units which influence the entire cross-section, that is all other units.
- In factor models they can often be modeled as observed common factors (Brownlees and Mesters, 2021).
- The degree a cross-sectional unit influences others varies.
- If a unit affects only the units closest to it, a shock of such unit will wear out when travelling through the network.
- This concept is called weak or spatial dependence and usually estimated by spatial methods.

Spatial dependence & Dominant units





Literature

- The identification of dominant units has recently received much attention
- Pesaran & Yang (2020), Brownlees, & Mesters (2021), Kapetanios et al. (2021) or Ditzen & Ravazzolo (2022).
- This paper follows the approach in Ditzen & Ravazzolo (2022). The authors suggest identifying dominant units using a two-step approach.
- In the first step a graphical network is estimated using a lasso estimator. (Meinshausen & Bühlmann, 2006; Sulaimanov & Koeppl, 2016).
- Second step uses column norms of the estimated network matrix to identify dominant units.
- Advantage of the approach is: robust to time dependence, common factors and heteroskedasticity.

Rigorous Lasso

- Ditzen & Ravazzolo (2022) find that the rigorous (or plugin) lasso (Bickel et al., 2009; Belloni et al., 2016; Ahrens et al., 2020) or the adaptive lasso (Zou, 2006; Medeiros & Menedes, 2016) works best to uncover the graphical representation in a framework with dominant units.
- Rigorous lasso uses a data-dependent, theory-driven penalization implementing a version of the lasso that allows for heteroskedastic and clustered errors; see Belloni et al. (2012, 2016).
- This makes it the first choice for empirical applications.

Sequentional Lasso estimator

$$\min_{\kappa_i} \frac{1}{T} \sum_{t=1}^{T} \left(x_{i,t} - x_{-i,t}\kappa_i'\right)^2 + \frac{\lambda}{T} \sum_{j}^{N} \psi_j \left|\kappa_j\right|$$

- $x_{i,t}$ is a $\mathrm{T} \times 1$ matrix containing the observations for the i-th unit. $x_{-i,t}$ is a $\mathrm{T} \times (\mathrm{N}-1)$ matrix containing all other cross-sections.
- κ_i is a $1 \times (\mathrm{N}-1)$ sparse vector containing past lasso OLS estimates.
- $\lambda = 2c\sqrt{T}\Phi^{-1}(1-\gamma/(2N))$. Commonly $c = 1.1, \gamma = 0.1/\log(T)$ and ψ_j depending on the explanatory variables and residuals is estimated.
- Ahrens, Aitken, Ditzen, Ersoy, Kohns and Schaffer (2020) show that ψ_j can be set such it allows for autocorrelation and heteroskedasticity.

$N \times N$

The estimated $\hat{\boldsymbol{\kappa}}_i$ are then stacked together $\hat{\boldsymbol{\kappa}} = (\hat{\boldsymbol{\kappa}}_1, \hat{\boldsymbol{\kappa}}_2, \ldots, \hat{\boldsymbol{\kappa}}_N)'$ into a $\mathrm{N} \times \mathrm{N}$ matrix, where the diagonal elements are zero:

$$\boldsymbol{\kappa} = (\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_{77}) = \begin{pmatrix} 0 & \kappa_{1,1} & \cdots & \cdots & \kappa_{1,77} \\ \kappa_{2,1} & 0 & \kappa_{2,3} & \cdots & \kappa_{2,77} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \kappa_{77,1} & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

Non-zero elements in $\boldsymbol{\kappa}_i$ imply that the respective cross-section influences crosssection i. In a graphical setting, the $\kappa_{ij} \neq 0$ implies that unit i and j are connected and thus they have an edge connecting them.

Identfying dominant units

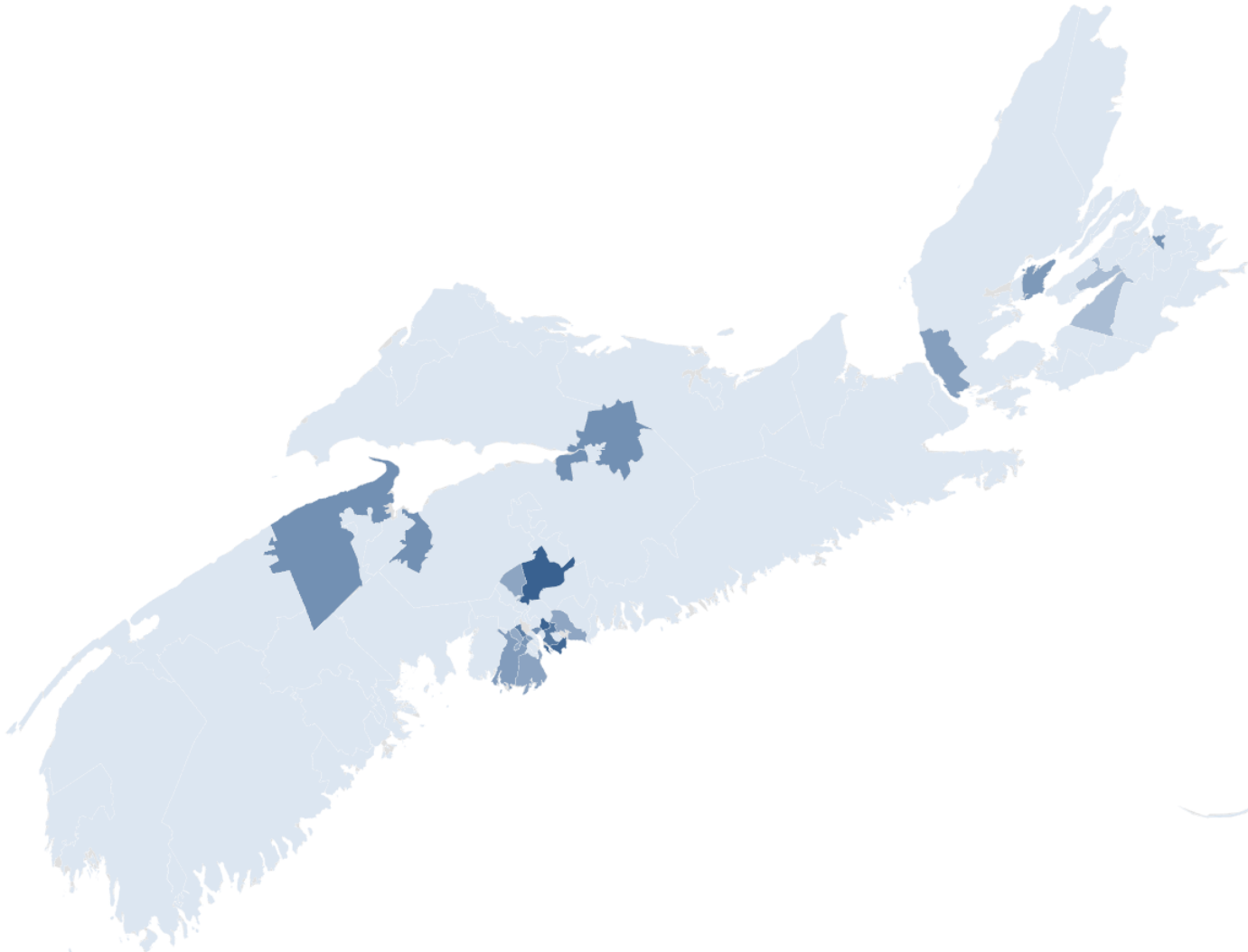
- Estimated coefficients are stacked into $\kappa = (\kappa_1, \ldots, \kappa_N)'$ and $\tilde{\kappa}_i$ is the i-th column of κ.
- Dominant units are defined in terms of their column norms, $\|\tilde{\boldsymbol{\kappa}}_i\|$, loosely following Ahn and Horenstein (2013) and Brownlees and Mesters (2021).
- A unit is a "global dominant unit" if the column norm is above a threshold $0 < c$ or among the largest k column norms.
- How do we define k? Following Brownlees and Mesters (2021) as:

$$k = \arg \max_{i=1,..,N} \|\tilde{\boldsymbol{\kappa}}_i\|_1 \, / \, \|\tilde{\boldsymbol{\kappa}}_{i+1}\|_1$$

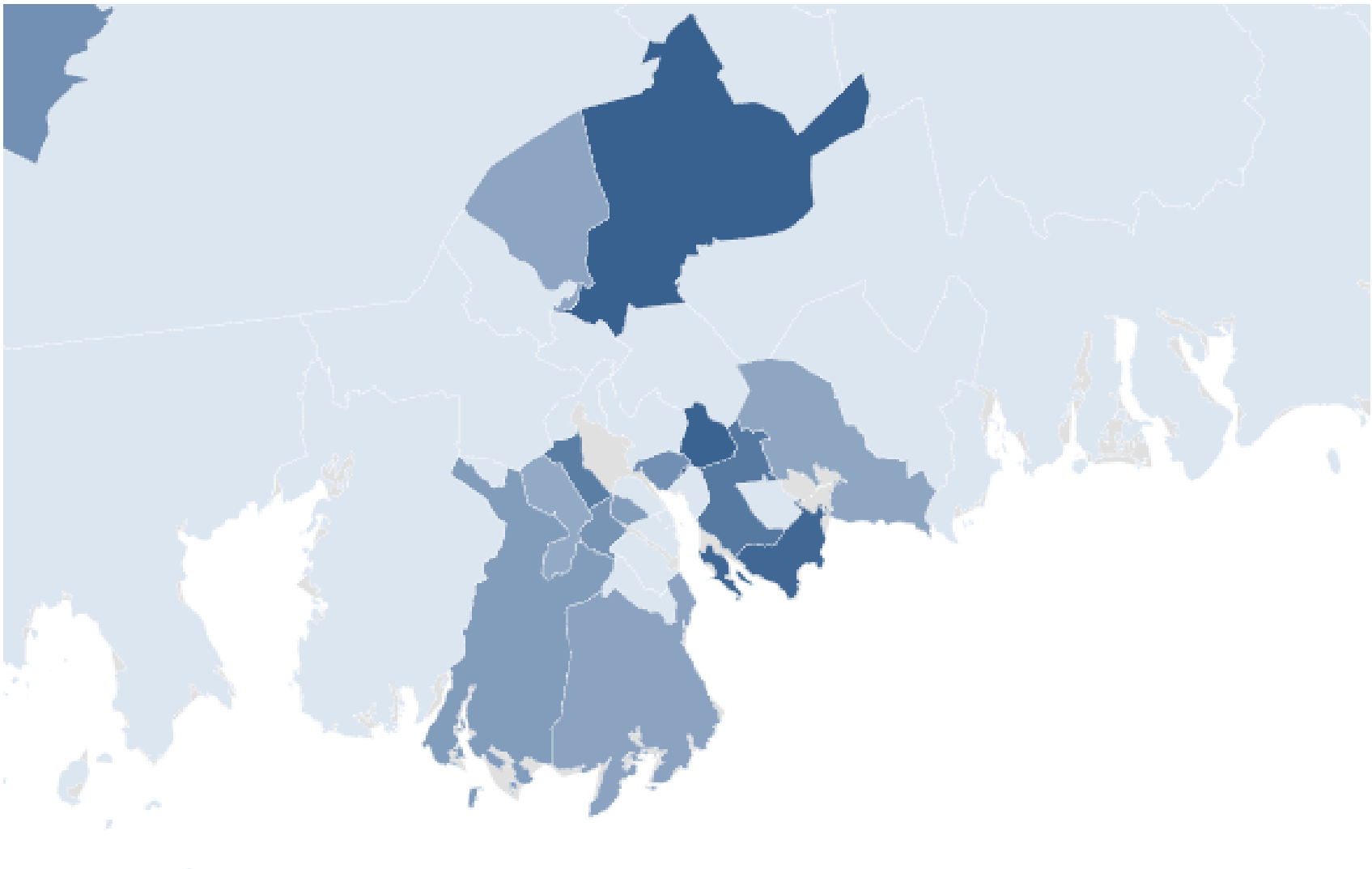
where the units are ordered following the column norms.

PS: Disadvantage is the assumption that there is at least one common factor.

We identify 18 “spreaders” out of 74 FSA’s

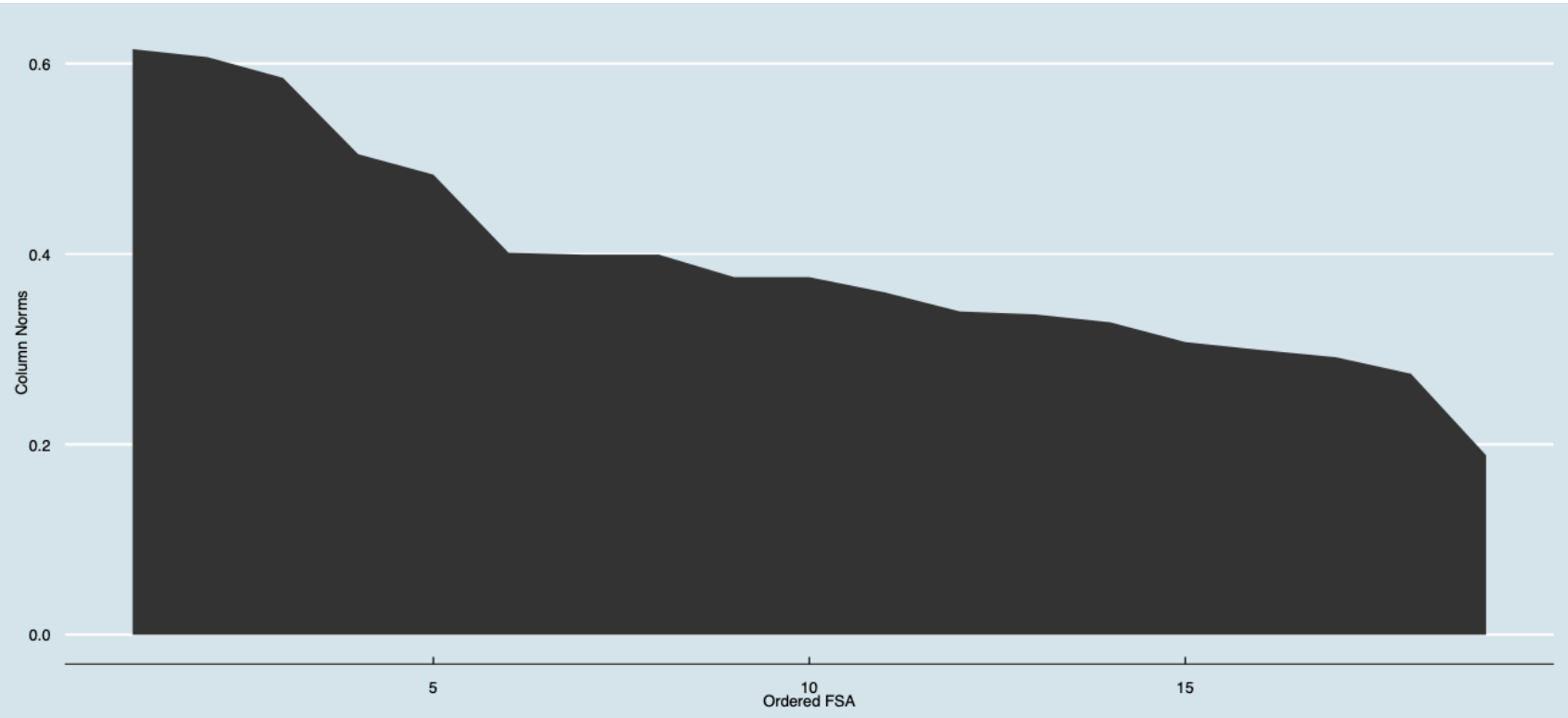


HRM

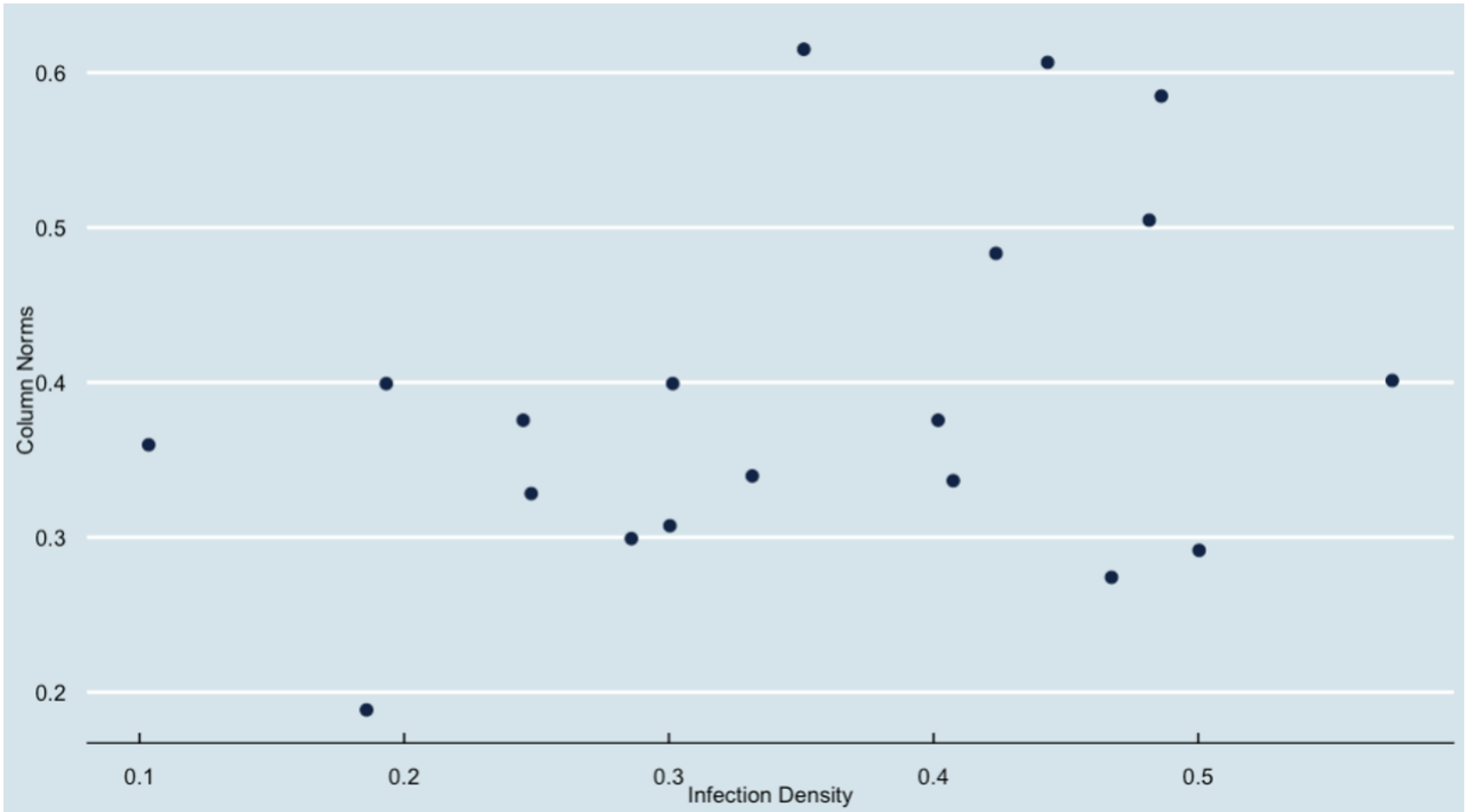


18 Spreaders, how?

The distribution of column norms ordered by their magnitudes



Column Norms vs. Infection Density



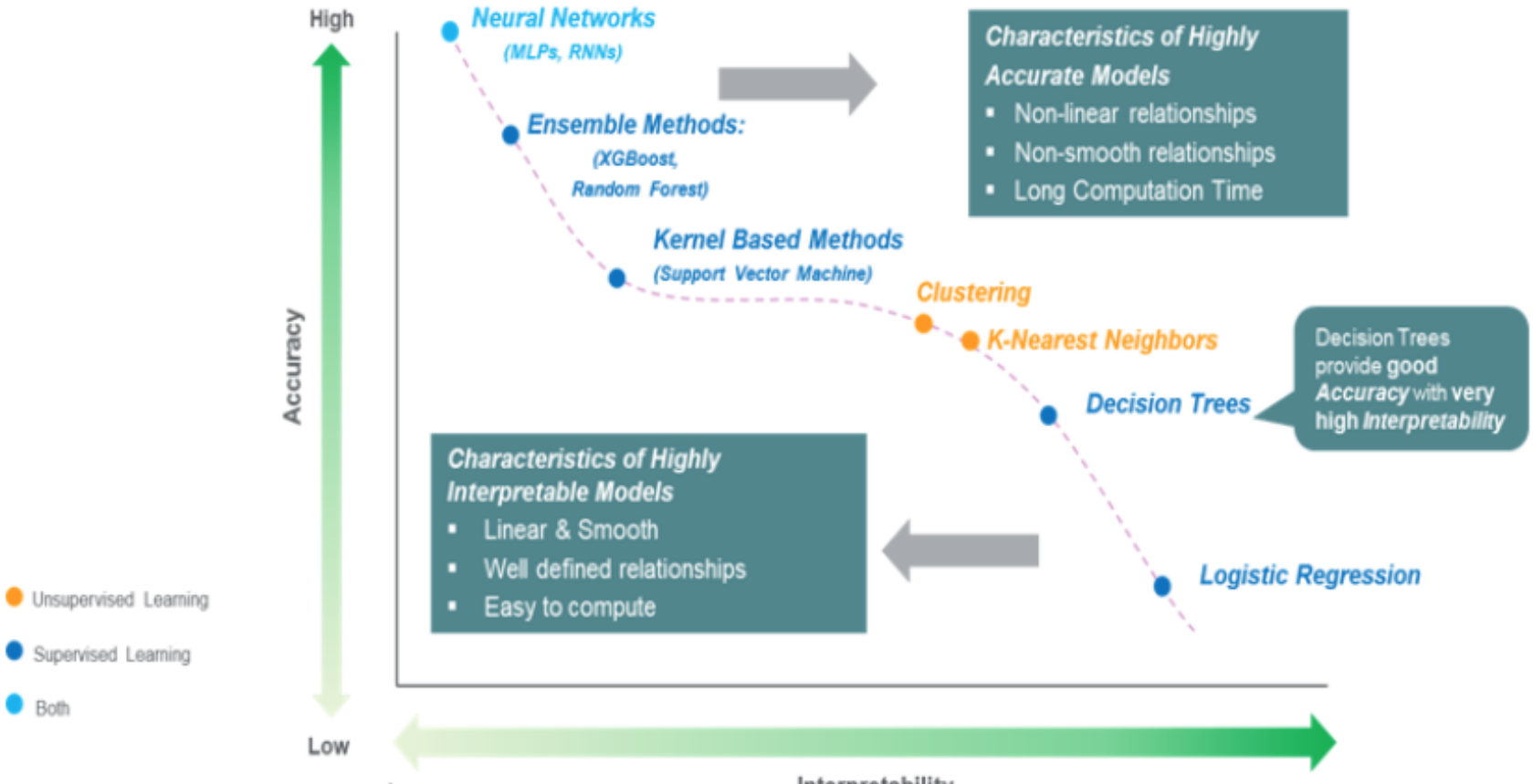
The correlation between column norms and infection densities is 0.31. The R-squared in the regression between norms and densities is 0.0807 with a statistically insignificant coefficient, 0.328.

What now?

- This shows that **a region with a low infection density could be a spreader** in the spatial network, while **a hotspot with a higher density could be a submissive region** in a viral spread.
- The important distinction in recognizing the regional spreaders is that the dominant unit analysis encompasses the temporal dynamics of the viral spread in its network structure, while the infections densities are just cumulative sums reflecting only cross-sectional differences.
- Next, the task is to see what regional characteristics make those eighteen regional spreaders different than the rest of the regions.

Important Predictors

- A ML model that accurately predicts outcomes is great, but most of the time you don't just need predictions, you want to be able to interpret your model.
- For example, if you build a model of house prices, knowing which features are most predictive of price tells us which features people are willing to pay for.



Choice is RF

- Random forests have been receiving increased attention as a means of variable selection in many classification tasks in bioinformatics and related scientific fields
- For instance, to select a subset of genetic markers relevant for the prediction of a certain disease.

Random Forest:

- It was introduced by Leo Breiman in 2001 and can be considered as an ensemble method
- It combines a large collection of trees.
- Each tree is based on selected bootstrap samples from observations (training) **and features**.
- Final predictions are obtained by voting all the trees.

Variable Importance (VI)

- Feature importance is the most useful interpretation tool to identify important association between the predicted outcome (cancer) and the features (genes).
- The most reliable method for VI is the "Mean Decrease in Accuracy".
- The more accurate our model (in prediction), the more we can trust the importance measures and other interpretations.

Problems with VI

- VI will be good only under some conditions
- Imagine that two variables can hold some redundant information by having a Spearman correlation $= 1$.
- These two variables will be interchangeable during growing the (Random) forest. Thus an equal amount of splits will rely on these two variables.
- The VI of both variables are the same. But if growing a new forest with one of the redundant variables omitted, the prediction performance could be almost unchanged, whereas the VI of the remaining redundant variable would double(if no other redundancies).
- Variables can also be complimentary or even interdependent. Hence, omitting an "important" variable would make the other variable less important.
- When a RF model essentially have captured a strong pair-wise variable interaction, VI can understate the loss of prediction performance by omitting one of the variables.

Solution: Unbiased Recursive Partitioning (URP)

- A Conditional Inference Framework (Hothorn et al. 2006).
- It avoids (according to authors) the variable selection bias of a tree based methods
- Unlike the others, URP uses a significance test procedure in order to select variables instead of selecting the variable that maximizes an information measure (e.g. Gini coefficient).
- The main difference seems to be that URP uses a covariate selection scheme that is based on statistical theory (i.e. selection by permutation-based significance tests) and thereby avoids a potential bias in `rpart`.
- Otherwise they seem similar; e.g. conditional inference trees can be used as base learners for Random Forests.

Application with 2016 Census

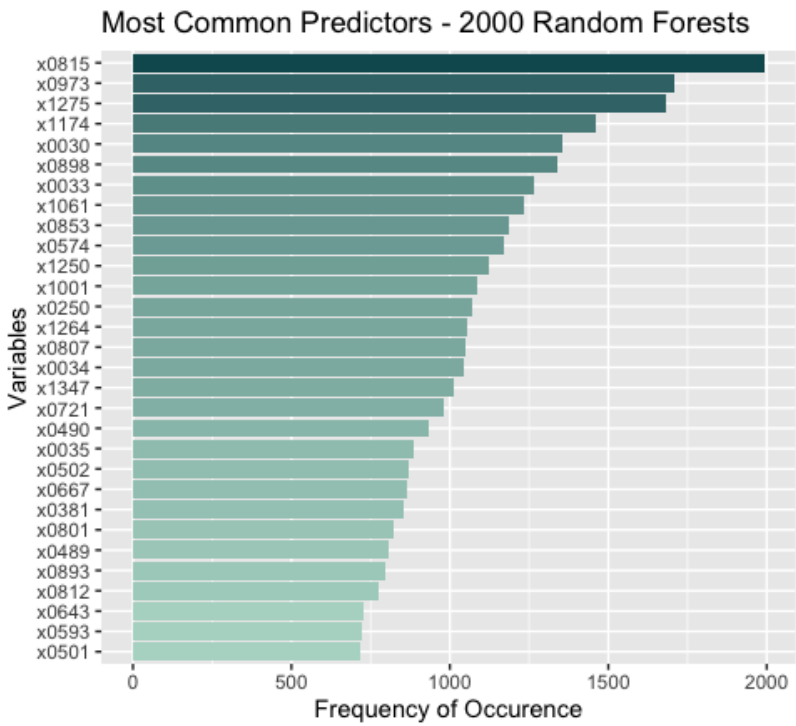
- The data for the socio-spatial risk factors are obtained from Canadian Census Analyser (CHASS) for the year 2016 at the FSA level.
- The census profile variables are grouped in 16 subcategories: Population and Dwellings, Age and Sex, Dwelling (dwelling characteristics and household size), Marital Status, Language, Income, Knowledge of Language, Immigration, Aboriginals and Visible Minorities, Housing, Ethnic Origin, Education, Labour, Journey to Work, Language of Work, Mobility.
- In each category, variables represent averaged values at each FSA and for each gender type.
- When we include all categories, we obtain more than 1400 socio-spatial variables for each of 74 FSA's in Nova Scotia.

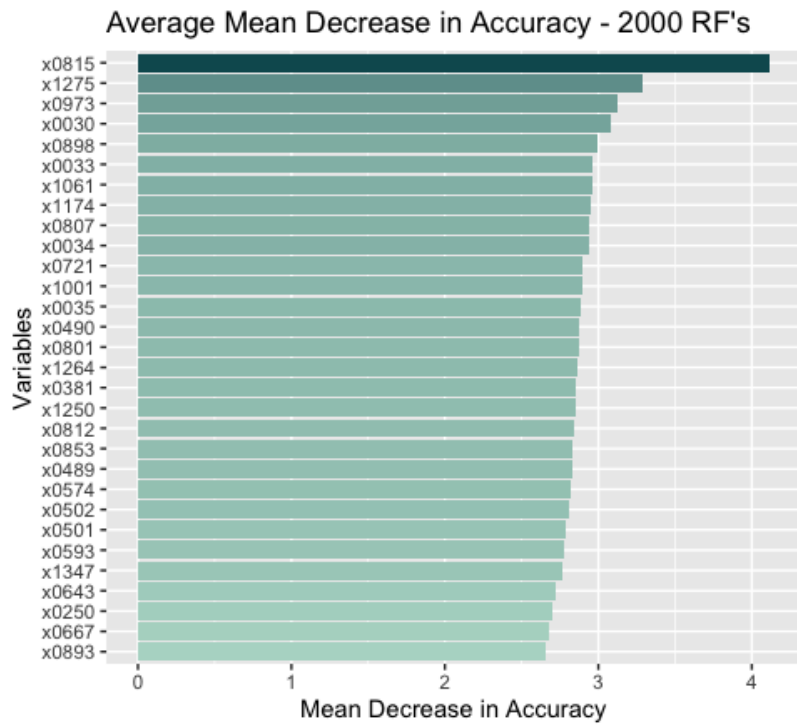
Classification with RF

- Initial data: 74 x 1378.
- Pre-processing: (1) zero and near zero variance variables are removed; (2) variables that are highly correlated with others (0.9 or above) are removed.
- We have 18 (1) "spreaders" and 58 (0) "followers".
- 2000 runs of RF application after tuning 2 hyperparameters
- Mean test AUC is 69.7% with a 95% CI 68.2%-70.8%
- We averaged 2000 VI graphs: Frequent Features and Top Features with permutation test

Socio-Spatial predictors of "Spreaders"

Conditional variable importance for random forests Strobl et al., 2008





Selective top predictors of "Spreaders"

- % of Acadian origin
- Higher LFP
- % of North African origin
- Lower average age
- Higher % of working population 15-65
- Higher % of military personnel
- Higher median rental payments
- Lower % of 65+
- Higher % of Inuit origin
- Higher % of South African origin
- Higher % of single income earners
- Higher median mortgage payments
- Higher % of foreign educated
- Higher % of Employment income
- Higher % of immigrants - PoB: Africa
- Higher % of people (journey to work) 60 min. +
- Higher % of recent immigrants - PoB: Africa
- Higher % of Mots Spoken language at home: non-official lang.

Concluding remarks

- First study that applies the networks analysis used in economics to a viral spread ...
- Keep in mind that predictors are as good as the predictive accuracy: 70%
- TBC: More exploratory analysis with Partial Dependence, SHAP (SHapley Additive exPlanations), Local Surrogate (LIME), and so on ...
- Plus some robustness checks ...

Thanks ...

- yigit.aydede@smu.ca
- Jan.Ditzen@unibz.it