UNIVERSITY OF ALBERTA
FACULTY OF SCIENCE
Department of Computing Science

UNIVERSITY OF ALBERTA
FACULTY OF MEDICINE & DENTISTRY

VizAR: A Software Tool for Epidemiological Hypothesis Generation with Geo-Spatial Data Mining

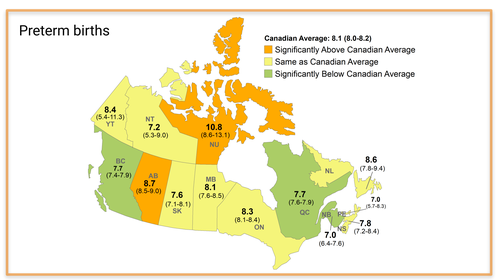
Bellinger C[1], Jabbar MSM[1], Hojjati S[1], Zaïane O[1], Osornio-Vargas A[2], and the DoMiNO Team

1. Computing Sciences, University of Alberta 2. Pediatrics, University of Alberta

Introduction

DoMiNo is an interdisciplinary research collaboration between epidemiologists, neonatologists, computer scientists, and public health officials.

The team includes combined expertise in paediatrics, epidemiology, geographical information systems, statistics and data mining.



The data mining team supports the development of hypotheses about the relationship between mixtures of airborne chemicals emitted from industry and adverse birth outcomes (ABOs) using geospatial data mining.

Objective

Design and develop a software tool to visualize association rules (VizAR) for use by researchers in public health, paediatrics and epidemiology, along with policy makers, that enables them to:

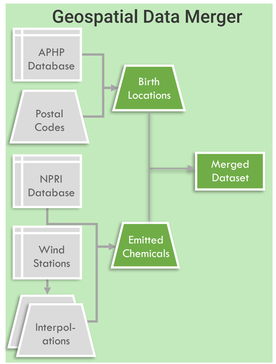
- Explore associations between airborne chemical mixtures and adverse birth outcomes discovered via geospatial data mining, and
- Identify associations relevant to their research for use in building new hypotheses for exploration.

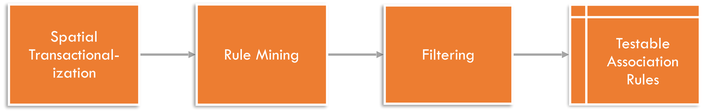
Methodology

Data: This research involved data from multiple sources including:

- ***Alberta Perinatal Health Program*** (APHP) dataset with 333,247 births, and 4 birth outcomes from 2006-2012, with valid maternal postal code locations
- ***National Pollutant Release Inventory*** (NPRI) database with annual reported point source emissions 136 chemicals from 2006-2012
- ***Alberta Agriculture weather station*** (AAWS) wind speed and direction

These were cleaned and integrated to form a merged geospatial dataset.





Data Mining: The geospatial dataset was transformed to a transactional dataset via `transactionalization' in order to facilitate the use of the kingfisher algorithm for data mining.

Our data mining process identified 1,700 association rules of interest (p-value<0.05) from over 1.6 million possible associations between chemicals and ABOs:

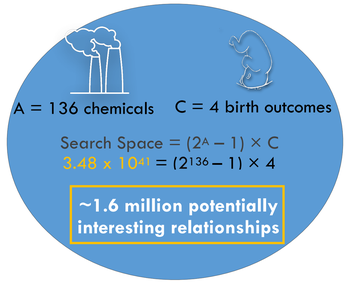
- chemical_1, chemical_2,… → ABO
- *e.g.* (PM, Methyl ethyl keytone, Toluene) → LBW

Software Development Process: Each researcher is interested in a unique subset of the association rules discovered.

We developed a visualization framework for association rules (VizAR) in order to facilitate exploration and the identification of new hypotheses.

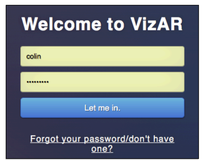
Our interdisciplinary team established a set of functional and non-functional requirements to support hypothesis generation with VizAR.

We employed an iterative process of software development, and user feedback was applied to ensure requirements were met.



Visualization Framework for Association Rules

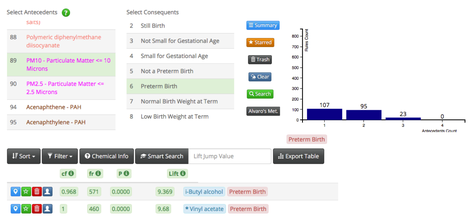
The VizAR software includes ***three levels of abstraction*** to enable users to discover association rules of interest at different levels of granularity.



Level 1: Instance level search and visualization

Users can ***search, sort and filter*** the discovered association rules to locate association involving chemicals and birth outcomes related to their research.

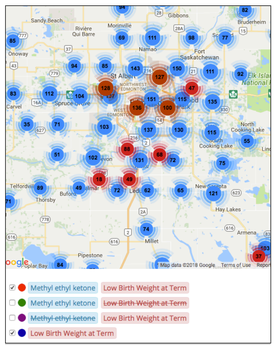
search, sort and filter



Once a rule of interest is identified, the ***geospatial distribution*** of the rule can be ***visualized*** in a Google Map.

- Visualization enables users to understand the geographic relationship between birth outcomes and industrial emission sources

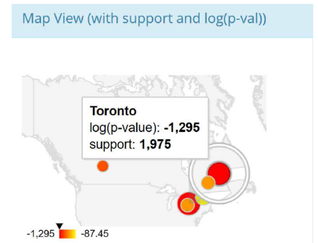
Visualize geospatial distribution



VizAR - Levels 2 and 3

Level 2: Pattern/Regional level

- Users can discover, regionally, common and contrasting association rules by selecting an individual rule and visualizing its occurrence across all regions



Level 3: High-level perspective

- Users can gain insights into the frequency and significance of rules in a region and across all regions by visualizing all association rules discovered across all regions.

Conclusions

The application of geospatial data mining to integrated multi-modal data:

- APHP, NPRI and AAWS

enabled the discovery of novel associations between chemical mixtures and adverse birth outcomes not possible with standard epidemiological methods.

VizAR enables efficient and effective knowledge transfer from the 1,700 associations rules discovered via geospatial data mining.

- VizAR facilitates users to explore the mined association rules and discover potential hypotheses to evaluate in their research.

References

1. Li J, Adilmagambetov A, Jabbar MSM, Zaiane O, Osornio-Vargas A, Wine O. On Discovering Co-Location Patterns in Datasets: A Case Study of Pollutants and Child Cancers. Geoinformatica. 20(4): 651-692, DOI 10.1007/s10707-016-0254-1, Apr 12, 2016
2. Jabbar MSM, Bellinger C, Zaiane O, Osornio-Vargas, A, "Discovering Co- location Patterns with Aggregated Spatial Transactions and Dependency Rules." International Journal of Data Science and Analytics
3. Jabbar MSM, Zaïane O, and Osornio-Vargas A. "Discovering spatial contrast and common sets with statistically significant co-location patterns." Proceedings of the Symposium on Applied Computing. ACM, 2017
4. Jabbar, MSM, and Zaïane O. "Learning statistically significant contrast sets." Canadian Conference on Artificial Intelligence. Springer, Cham, 2016.
5. Hämäläinen W. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. Knowledge Information Systems. vol. 32, page 383–414, 2012

Acknowledgements

- Data – Alberta Perinatal Health Program (APHP) www.aphp.ca
  National Pollutant Release Inventory (NPRI) www.ec.gc.ca/inrp-npri
- Team – Data Mining and Neonatal Outcomes (DoMiNO) Project
- Funding – CIHR/NSERC Collaborative Health Research Program (2013-2016)

Contact

e: cbelling@ualberta.ca
w: www.ualberta.ca/~cbelling
w: https://sites.google.com/a/ualberta.ca/domino/