Environment International 131 (2019) 104972

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Interdisciplinary-driven hypotheses on spatial associations of mixtures of industrial air pollutants with adverse birth outcomes

Jesus Serrano-Lomelin[a,h], Charlene C. Nielsen[b,c], M. Shazan M. Jabbar[d], Osnat Wine[b], Colin Bellinger[d], Paul J. Villeneuve[e], Dave Stieb[f], Nancy Aelicks[g], Khalid Aziz[b], Irena Buka[b], Sue Chandra[h], Susan Crawford[g], Paul Demers[i], Anders C. Erickson[j], Perry Hystad[k], Manoj Kumar[b], Erica Phipps[l], Prakesh S. Shah[m], Yan Yuan[a], Osmar R. Zaiane[d], Alvaro R. Osornio-Vargas[b,*]

[a] School of Public Health, University of Alberta, Edmonton Clinic Health Academy, 11405 87 Avenue, Edmonton, Alberta T6G 1C9, Canada
[b] Department of Pediatrics, University of Alberta, Edmonton Clinic Health Academy, 11405 87 Avenue, Edmonton, Alberta T6G 1C9, Canada
[c] Department of Earth and Atmospheric Sciences, University of Alberta, 1-26 Earth Science Building, Edmonton, Alberta T6G 2E3, Canada
[d] Department of Computing Science, University of Alberta, 32 Athabasca Hall, Edmonton, Alberta T6G 2E8, Canada
[e] Department of Health Sciences, Carleton University, Herzberg Building, Room 5413, 1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada
[f] Environmental Health Science and Research Bureau, Health Canada, 50 Colombine Driveway, Ottawa, Ontario K1A 0K9, Canada
[g] Alberta Health Services, Alberta Perinatal Health Program, Suite 310, 1403-29 Street NW, Calgary, Alberta T2N 2T9, Canada
[h] Department of Obstetrics & Gynecology, University of Alberta, Royal Alexandra Hospital, 10240 Kingsway Avenue, Edmonton, Alberta T5H 3V9, Canada
[i] CAREX Canada, Faculty of Health Sciences, Simon Fraser University, 105-515 West Hastings St, Vancouver, BC V6B 5K3, Canada
[j] School of Population and Public Health, University of British Columbia, 2206 E Mall, Vancouver, BC V6T 1Z3, Canada
[k] School of Biological and Population Health Sciences, Oregon State University, 101 Milam Hall, Corvallis, OR 97331, USA
[l] Canadian Partnership for Children's Health & Environment, 1500-55 University Avenue, Toronto, Ontario M5J 2H7, Canada
[m] Department of Pediatrics and Institute of Health Policy, Management, and Evaluation, University of Toronto, Mount Sinai Hospital, 600 University Avenue, Room 19-231A, Toronto, Ontario M5G 1X5, Canada

ARTICLE INFO

Handling Editor: Hanna Boogaard

Keywords:
Adverse birth outcomes
Air pollution
Industrial air emissions
Spatial data mining
BTEX group
Particulate matter

ABSTRACT

Background: Adverse birth outcomes (ABO) such as prematurity and small for gestational age confer a high risk of mortality and morbidity. ABO have been linked to air pollution; however, relationships with mixtures of industrial emissions are poorly understood. The exploration of relationships between ABO and mixtures is complex when hundreds of chemicals are analyzed simultaneously, requiring the use of novel approaches.

Objective: We aimed to generate robust hypotheses spatially linking mixtures and the occurrence of ABO using a spatial data mining algorithm and subsequent geographical and statistical analysis. The spatial data mining approach aimed to reduce data dimensionality and efficiently identify spatial associations between multiple chemicals and ABO.

Methods: We discovered co-location patterns of mixtures and ABO in Alberta, Canada (2006–2012). An ad-hoc spatial data mining algorithm allowed the extraction of *primary* co-location *patterns* of 136 chemicals released into the air by 6279 industrial facilities (National Pollutant Release Inventory), wind-patterns from 182 stations, and 333,247 singleton live births at the maternal postal code at delivery (Alberta Perinatal Health Program), from which we identified cases of preterm birth, small for gestational age, and low birth weight at term. We selected *secondary patterns* using a lift ratio metric from ABO and non-ABO impacted by the same mixture. The relevance of the *secondary patterns* was estimated using logistic models (adjusted by socioeconomic status and ABO-related maternal factors) and a geographic-based assignment of maternal exposure to the mixtures as calculated by kernel density.

Results: From 136 chemicals and three ABO, spatial data mining identified 1700 *primary patterns* from which five

* Corresponding author at: University of Alberta, Department of Pediatrics, 3-591 ECHA, 11405 87th Avenue, Edmonton, Alberta T6G 1C9, Canada.
E-mail addresses: jaserran@ualberta.ca (J. Serrano-Lomelin), ccn@ualberta.ca (C.C. Nielsen), mohomedj@ualberta.ca (M.S.M. Jabbar), osnat@ualberta.ca (O. Wine), cbelling@ualberta.ca (C. Bellinger), Paul.Villeneuve@carleton.ca (P.J. Villeneuve), Dave.Stieb@hc-sc.gc.ca (D. Stieb), nancy.aelicks@albertahealthservices.ca (N. Aelicks), khalid.aziz@ualberta.ca (K. Aziz), ibuka@ualberta.ca (I. Buka), Sue.Chandra@albertahealthservices.ca (S. Chandra), susan.crawford@albertahealthservices.ca (S. Crawford), Paul.Demers@cancercare.on.ca (P. Demers), anderse@uvic.ca (A.C. Erickson), manojk@ualberta.ca (M. Kumar), erica@healthyenvironmentforkids.ca (E. Phipps), pshah@mtsinai.on.ca (P.S. Shah), yyuan@ualberta.ca (Y. Yuan), zaiane@cs.ualberta.ca (O.R. Zaiane), osornio@ualberta.ca (A.R. Osornio-Vargas).

https://doi.org/10.1016/j.envint.2019.104972
Received 22 March 2019; Received in revised form 26 June 2019; Accepted 26 June 2019

secondary patterns of three-chemical mixtures, including particulate matter, methyl-ethyl-ketone, xylene, carbon monoxide, 2-butoxyethanol, and n-butyl alcohol, were subsequently analyzed. The significance of the associations (odds ratio > 1) between the five *mixtures* and ABO provided statistical support for a new set of hypotheses.

Conclusion: This study demonstrated that, in complex research settings, spatial data mining followed by pattern selection and geographic and statistical analyses can catalyze future research on associations between air pollutant mixtures and adverse birth outcomes.

1. Introduction

There is much interest in methods for understanding the adverse impacts of exposure to mixtures of air pollutants on human health since ambient air pollution is composed of mixtures of chemicals (Dominici et al., 2010, Mauderly et al., 2010, SCHER, 2012). Recent evidence suggests that mixtures of chemicals can have a toxicological behavior that differs from the toxicity of the individual chemicals. For example, the effect of ozone (O_3) on asthma may be higher when it coexists with other co-pollutants such as sulphur dioxide (SO_2) and particulate matter (PM) than the effect of ozone alone (Toti et al., 2016). Likewise, clusters of elevated NO_2, NO, and $PM_{2.5}$ concentrations may increase the odds of low birth weight relative to the effects of each one separately (Coker et al., 2016). Thus, findings from mixture studies could guide new environmental policies based on a multipollutant framework to mitigate exposure to chemical emissions (Hidy and Pennell, 2010).

Studies considering mixtures of multiple pollutants are difficult to implement due to a variety of potential limitations. Limitations include the lack of monitored data for many chemicals, the limited toxicological information for many of the chemicals produced/emitted worldwide (SCHER, 2012), difficulties in assessing spatiotemporal variation of groups of interacting pollutants in the atmosphere to assess human exposure, and the lack of statistical methods to parsimoniously assess effects of pollutant mixtures (Mauderly et al., 2010).

Advances in multipollutant approaches in relation to the adverse health effects of air pollution have been carried out in urban settings, where a small number of pollutants are regularly monitored (e.g., criteria air pollutants) (Edwards et al., 2015; Wilhelm et al., 2012). In contrast, few studies have considered alternative sources of data for a broader set of chemicals emitted by industrial facilities into the environment (Currie and Schmieder, 2009, Agarwal et al., 2010, Wine et al., 2014, Willis and Hystad, 2019). The large number of possible combinations or mixtures is an issue of methodological and theoretical concern, especially when the toxicity of participating chemicals is unknown (SCHER, 2012). Expanding the number of chemicals increases the complexity for identifying hazardous mixtures when no specific hypothesis drives a study, thus motivating researchers to use novel methodological approaches as an initial filter of potential mixtures of health concern.

Recently, data mining algorithms have been used to rapidly scan large numbers of pollutants and filter the ones that may require a more formal future study. Frequent itemset mining algorithms were used to identify common chemical combinations in human populations from a set of 106 chemicals (Kapraun et al., 2017) and to estimate relationships between chemicals and health biomarkers of diseases (Bell and Edwards, 2015). Association rule mining algorithms were used to identify patterns of air pollutant-combinations related to pediatric asthma exacerbations (Toti et al., 2016), as well as to find significant spatial co-location patterns of childhood cancer occurrence and chemicals released into the air by industrial sources in Canada (Li et al., 2016). More recently, Jabbar et al. (2018) used and modified the algorithm developed by Li et al. (2014) to identify statistically significant spatial co-location patterns of adverse birth outcomes and mixtures of industrial chemicals in Canada. They presented hundreds of statistically significant patterns, suggesting that post-selection tasks are necessary to select those mixtures of potentially higher health concern.

Adverse birth outcomes (ABO) are an essential health outcome to assess due to their implications for human development and health outcomes throughout the lifespan as prioritized by the World Health Organization under the Sustainable Development Goals (World Health Organization, 2018). Ample epidemiological research indicates that babies born before completing their gestational period (preterm birth: PTB), weighing < 2.5 kg at 37 or more weeks of gestation (low birth weight at term: LBWT), or weighing less than expected for their sex and gestational age (small for gestational age: SGA), have lower chances of survival and higher probabilities of developing chronic diseases throughout life (Kramer et al., 2001; Kramer, 2003).

A variety of individual (Heaman et al., 2013), social (Kim and Saada, 2013, Auger et al., 2009), and environmental factors, including air pollution (Stieb et al., 2012, 2015), have been recognized as risk factors for ABO. The effects of air pollutants on fetal development throughout the pregnancy period are increasingly studied (e.g., Wang et al., 2018) since fetal susceptibility to environmental issues varies over the gestational period.

Although the research on air pollution and perinatal epidemiology has been extensive, it has mainly focused on studying single pollutants primarily from traffic sources (Wigle et al., 2008; Stieb et al., 2016). Consequently, the effects of air pollutant mixtures from industrial activities on ABO are poorly understood (Slama et al., 2008). Some related studies have followed approaches based on the proximity to specific sources of industrial pollutants (e.g. Walker Whitworth et al., 2018, Casey et al., 2016). However, a particular concern is to understand the potential consequences of being exposed to a disproportionate number of recognized hazardous chemicals from different industrial sources during pregnancy and early development (Slama et al., 2008; Giudice, 2016; Sutton et al., 2012; Wang et al., 2016). Understanding the role of a large number of chemicals that have not been extensively explored may shed light on lesser-known chemicals and possible relationships with ABO. Addressing this gap is challenging due to the potentially large number of chemical mixtures that could be present in ambient air, usually where industrial activity is pervasive. In this regard, the Scientific Committee on Health and Environmental Risks of the European Union has suggested to use some form of initial filter to allow a focus on mixtures of potential concern (SCHER, 2012). No specific individual chemicals need to be the target of such a study, but rather all possible industrial chemicals that have not yet been suspected or hypothesized of having associations with ABO may be more efficiently identified.

Therefore, we aimed to extract candidate hypotheses linking ABO with mixtures originating from hundreds of chemicals released by industry into the air. We used a pruning approach for the spatial data mining algorithm developed by Jabbar et al. (2018) to identify and prioritize candidate hypotheses in complex settings. Subsequent spatial and regression analysis were then conducted to assign statistical robustness to those hypotheses. This systematic approach could guide future research on specific mixtures related to ABO.

2. Methods

2.1. Research framework

This study was part of a national research project on Data Mining

and Neonatal Outcomes (DoMiNO) described elsewhere (Wine et al., 2019). The primary objective of DoMiNO was to identify spatial co-location of ABO with multiple combinations of industrial chemicals emitted to air in Canada, of which the results may serve as a foundation for future research. The project was built on an ongoing collaboration among specialists in perinatology, neonatology, computer-sciences, geography, exposure science, epidemiology, and knowledge users from government, data-provider agencies, and non-governmental organizations. A dedicated interdisciplinary collaborative research approach (i.e. integrated knowledge translation) (CIHR, 2012) supported the progression of the different phases of this project.

It is essential to state that the scope of DoMiNO was exploratory. Causal relationships between mixtures of chemicals and ABOs were not expected. The characteristics of our data sources limit our capacity to go beyond an exploratory analysis. For example, our source of industrial chemicals (the National Pollutant Release Inventory) contains only annual estimations of chemicals emitted into the environment (air, water, soil) by industrial facilities (see Section 2.4) reducing the ability to untangle temporary effects of mixtures throughout the pregnancy period and forcing us to simplify exposure assignment. Assumptions, simplifications, and limitations are explained across the manuscript.

The study received approval from the University of Alberta Health

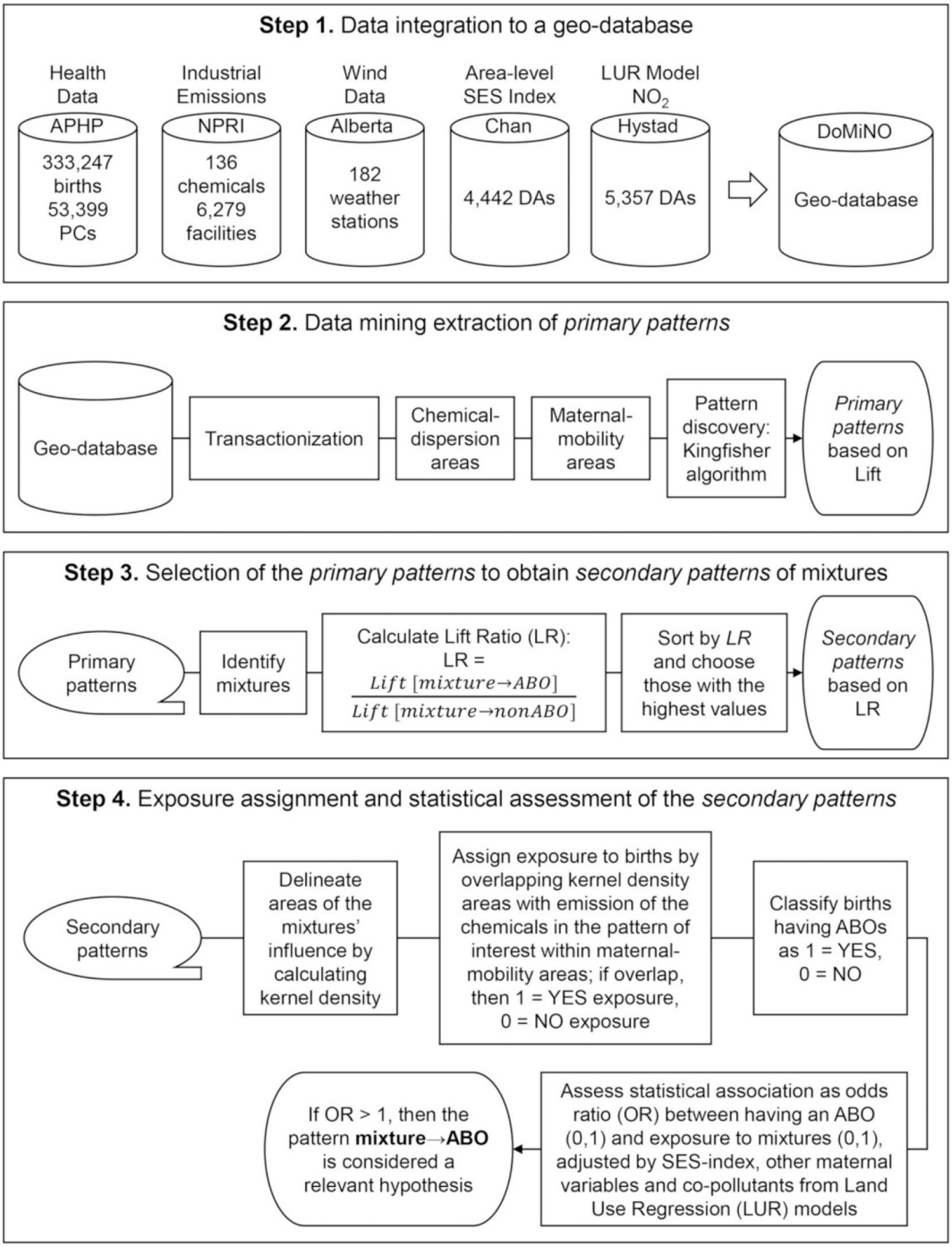

Fig. 1. The four steps of the interdisciplinary framework involved spatial data mining, geographical information systems, and bio-statistics to integrate the health and environmental variables into a geodatabase based on postal codes (PC), extract primary patterns, select secondary patterns, and assess as relevant hypotheses for associations between mixtures of industrial chemical emissions and adverse birth outcomes (ABO) using odds ratios (OR).

Research Ethics Board Human Panel (Study ID Pro00039545) and from the Alberta Perinatal Health Program.

2.2. Study setting

This study focused on the province of Alberta, Canada. Alberta has had one of the highest rates of ABO in recent years (PHAC, 2013) and the highest number of emitting facilities in the country (Environment and Climate Change Canada, 2016). In addition, the province has high-quality health data.

According to Statistics Canada (2014), between 2005 and 2007 the prevalence of PTB, SGA, and LBW in Alberta (8.7%, 8.8%, and 6.7%, respectively) had been consistently higher than the national averages (7.8%, 8.4%, and 6.0%, respectively). Furthermore, these disorders related to short gestation and low birth weight are consistently ranked as the 2nd leading cause of infant mortality (Statistics Canada, 2012).

The project combined five large databases containing: birth outcomes, industrial emissions, wind speed and direction, area-level socioeconomic status, and area-level NO_2 concentrations (summarized in Supplemental Table S1). Multiple procedures were used to develop the hypothesis generation framework (outlined in Fig. 1). Datasets and integration procedures are described in detail below.

2.3. Health data

We included birth and maternal data from all singleton live births in Alberta that occurred between 2006 and 2012. Another set of birth data, from 2013 to 2014, was used for additional testing of the resulting hypotheses from the first data set. Data were extracted from the Alberta Perinatal Health Program database (Alberta Health Services, 2014), which covers the entire birth population of children delivered in hospitals, as well as planned home births, and unplanned deliveries outside a facility. Anonymized data from single live births (> 21 weeks of gestational age) were analyzed. The records included maternal age, maternal residence postal code (a six-character alphanumeric combination assigned to one or more postal addresses), newborns' birth date, birth weight, type of labor (spontaneous, induced), and gestational age at delivery in completed weeks.

In the Alberta Perinatal Health Program database, we identified cases of PTB (newborns at < 37 weeks of gestation), SGA (newborns with a birth weight below the tenth percentile weight from a sex-age specific Canadian-population reference: Kramer et al., 2001, Kramer, 2003), and LBWT (newborns weighing < 2500 g at 37 or more weeks).

The Alberta Perinatal Health Program included obstetrical, pre-pregnancy and during-pregnancy variables from medical records, which were identified as risk factors for ABO based on other studies (Tough et al., 2001; Serrano-Lomelin, 2017). The obstetrical factors included past-preterm, past-SGA, and parity. The pregnancy-related factors included mothers' weight < 45 kg and other medical disorders. The during-pregnancy factors included gestational hypertension, gestational diabetes, smoking, substance use, and bleeding anytime during pregnancy (before or after the 20th week). All definitions are included in the Supplemental Table S2.

2.4. Industrial emissions

Chemicals released to the air, as reported annually by industrial facilities, were extracted from the National Pollutant Release Inventory. The Canada-wide database included the releases (above specific-chemical thresholds) of 342 chemicals to air, water, and land, and the geolocation of the facilities (i.e. longitude and latitude coordinates) (Environment and Climate Change Canada, 2015). We selected only the chemicals released into the air in Alberta during 2006–2012 (n = 136), and for the subsequent assessment for the years 2013–2014. The emissions reported in kilograms (kg) and grams (g) were converted to tonnes. We calculated the average of each chemical emitted by each facility over the entire time periods for subsequent analysis.

2.5. Wind data

We acquired wind data from 182 stations in Alberta Agriculture's AgroClimatic Information System (ACIS, 2010) for 2006–2012. We calculated the mean wind speed overall seven years and interpolated the mean wind direction (Williams, 1999) using the spline function in ArcGIS Desktop (Esri, 2016). We interpolated raster surfaces from 156 points representing the trigonometric X and Y dimensions of the average wind direction angle using a two-dimensional curvature spline technique. Specifically, the regularized spline was based on the defaults 0.1 for the weight and 12 for the number of points. The mean values were assigned to the facility locations for input to the data mining algorithm.

2.6. Area-level socioeconomic status index

We used a small area-level Canadian socioeconomic status index (SES-index) developed by Chan et al. (2015). This index uses data from the 2006 National Census. It incorporates, in a single index, census data on education level, employment status, income, marital status, home ownership, transport mode, year of home-construction, and the aboriginal status or human developmental index of the individuals' country-origin, among other variables. Thus, the SES-index captures relevant information for ABO. The index value was originally assigned to each dissemination area and reported in quintiles. A low SES-index quintile indicates low socioeconomic status. Dissemination Areas (DA) are census geographic areas with a population of 400 to 700 persons (Statistic Canada, 2008), which group postal codes. The maternal postal codes at delivery were assigned to DA using boundary file identifiers (Statistic Canada, 2007) and a vector overlay (point in polygon). Each birth record was linked to the corresponding SES-index based on the geographic link between postal codes and DA.

2.7. Area-level concentrations of NO_2

We used area-level nitrogen dioxide (NO_2) concentrations derived from a national land use regression model (Hystad et al., 2011) as a surrogate for the contribution of other pollutant sources (e.g., traffic-related). NO_2 concentrations were used as a covariate in the statistical models (see Section 2.13). Variation in regional and local-scale pollution was captured in multiple regression models, which incorporated satellite-based estimates, fixed-site monitoring measurements, and geographic predictor variables for the year 2006. As done for the SES values, we used vector overlay to assign the dissemination area-level measures to the postal codes of the birth records.

2.8. Data analysis

The analysis consisted of four major steps: (1) data integration linking health outcomes and environmental data at the postal code level into a geodatabase; (2) extraction of significant spatial co-location patterns of chemicals and ABO or non-ABO cases, named *primary patterns*, using spatial data mining; (3) selection of *secondary patterns* for mixtures of chemicals using a pruning metric, and; (4) the exposure assignment and statistical assessment of the *secondary patterns* using GIS methods and regression analysis (Fig. 1).

2.9. Data integration

The geo-database consisted of the attributed maternal postal codes (the birth data table) and the National Pollutant Release Inventory locations (the facility locations table). For the birth data, each 6-character postal code of the maternal residence at the time of birth, from the Alberta Perinatal Health Program, was linked to the longitude and

latitude coordinates from Digital Mapping Technology Inc. (DMTI) Spatial's Postal Code Suite (DMTI, 2014). The birth data table contained the individual birth and maternal variables and merged with the area-level SES-index and NO_2 values at the postal code level. The facility locations table contained the chemical emission and wind variables. Those two tables were used, independently, for data mining and GIS and regression analysis.

2.10. Data mining extraction of primary patterns

We applied the spatial co-location pattern mining algorithm AGT-Fisher (Aggregated Grid Transactionalization in conjunction with the Fisher's test-based and Kingfisher dependency rule search technique; Jabbar et al., 2018). The primary goal of this spatial algorithm was to find relevant primary co-location patterns based on the spatial overlap of air pollutant emission regions and maternal mobility regions during pregnancy. Such patterns explained which individual or combinations of industrial air pollutants were co-located, or in near proximity, with live births, including ABO and non-ABO. The AGT-Fisher consisted of two major processes: grid transactionalization and pattern discovery.

The transactionalization process consisted of transforming spatial data into transactions (Li et al., 2014; Jabbar et al., 2018). For this, we overlaid the region of interest (map) with a set of uniformly distributed grid points (1-km grid). Each grid point recorded the occurrence or absence (binary true/false) of each event (ABO or non-ABO) and each industrial chemical at its location. Each grid point was added to the transactional database that was subsequently mined with the Kingfisher algorithm. An example grid point transaction is {SGA = True, LBW = False, …, benzene = True, chlorine = False, PM = True, …}. In such a dataset, each record was considered as a transaction where it contained a set of co-occurring events or items. For our research, the items consisted of overlapping regions. To determine the overlapping regions, we generated the dispersion region of an air pollutant from an emission point (facility) as a circular buffer where the center was the emission point, and the radius was defined based on the chemical's amount released. Then, we altered the circular region into an elliptical buffer region based on the period's average wind speed and direction to better approximate the actual chemical dispersion and the area of exposure to chemicals. The lengths of the major- and minor-axis (a and b, respectively) were computed as follows: $a = r + \gamma\,|\nu|$; $b = r^2/a$; where r was the radius of the initial circle, and it was equal to the natural logarithm of the amount of chemical released at a given location [$r = \ln(amounts)$]; ν was the wind speed, and γ was the stretching coefficient ($=0.3$). Detailed information about this process has been published by Jabbar et al. (2018). For the birth data, we generated mobility regions as 5 km radius circles centered on the postal code location of the maternal residence, as a surrogate of the maternal mobility range during pregnancy. Finally, by overlaying a 1-km grid, we computed the transactions required for identifying primary patterns (Fig. 2). The transformed dataset enabled us to utilize the Kingfisher algorithm (Hamalainen, 2012) to discover nonspurious co-location patterns. Our previous work (Jabbar et al., 2018) demonstrated that the Kingfisher algorithm was statistically efficient for finding non-redundant statistically significant co-location patterns between chemical mixtures and ABO. Kingfisher judged the statistical significance of the association between chemical mixtures and ABO using Fisher's exact test. The algorithm used enumeration trees to search and prune the co-location patterns, thereby discovering likely patterns in a computationally efficient manner. The AGT-Fisher algorithm discovered a set of co-location patterns of the form *chemical → ABO or chemical → non-ABO*, where the pattern satisfied a p-value threshold. We used a p-value cut-off of 0.05, which is a standard cut-off commonly used in data mining algorithms.

2.11. Secondary patterns of mixtures

Although AGT-Fisher narrowed the scope of the spatial associations between chemical mixtures and birth outcomes, in such a high-dimensional dataset the list of such associations was still very large (i.e., hundreds). Moreover, it was highly likely that only a small subset of these patterns would be of interest to knowledge users. So, in order to reduce the set of discovered patterns to a more focused interesting subset, we further utilized the standard data mining metric called lift (Park et al., 2014). Given a co-location pattern X → Y, lift measured the statistical dependency between the occurrences of X (i.e. chemical mixtures) and the occurrences of Y (i.e. ABO). In other words, the lift of the rule *chemical → ABO* related the probability of the chemical and the ABO occurring together to the probability of the chemical and the ABO occurring separately (independence condition). The lift of an association X → Y, *lift* (X, Y), ranges from zero to infinity, where a lift of 1 indicates independence and a lift greater than one indicates a positive spatial dependence (Park et al., 2014).

Since the intent was to identify co-location of a group of chemicals within the overlapping regions (called *mixtures* in our study), we focused on searching for patterns with the highest lifts that had two or more chemicals in the pattern. For those patterns, we calculated the lift ratio (LR), which is the ratio of the lift of exposure to the chemical mixture and having the ABO over the lift of exposure to the chemical mixture and not having the ABO (non-ABO): LR = {*lift* (*mixture* →

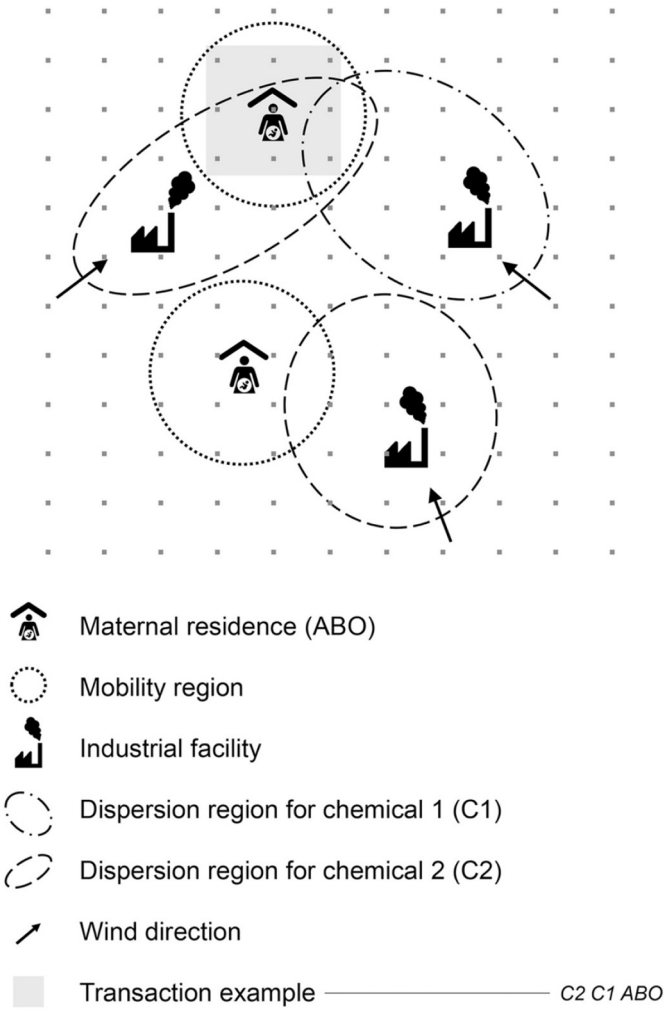

Fig. 2. The spatial data mining algorithm extended the maternal residences and chemical emission sources, and the overlap of these spatial objects were counted at grid points to create the transactions required to identify each statistically significant/prevalent association pattern where a pollutant mixture for two chemicals (chemical 1 = C1 and chemical 2 = C2) was co-located with an adverse birth outcome (ABO).

ABO) / *lift* (*mixture → non-ABO*)}.

That means LR takes the ABO occurrence and ABO nonoccurrence possibilities from the same chemical(s). By following this metric, we are assured that, under the data mining framework, the spatial dependency of the *mixture → ABO* is greater than the spatial dependency of the *mixture → non-ABO*. These secondary co-location patterns provide a list of strong potential hypotheses to be researched by domain experts, i.e. geographic information systems techniques and regression analysis for confirmation.

2.12. Assignment of spatial exposure-areas to secondary patterns

Since the spatial data mining algorithm worked with transactions instead of individuals (cases and no cases), we used geographic information systems (GIS) to delineate coexisting areas of mothers with live births and mixtures where exposure may have occurred. The area of potential exposure to industrial chemicals of those mothers during pregnancy was calculated by applying kernel density (Silverman, 1997). We estimated the spread of industrial emissions from their point sources as mean tonnes/km^2 using ArcGIS Desktop (Esri, 2016; Nielsen et al., 2017). We parameterized the kernel density maps for each chemical using a 1000-m cell size and a 10-km radius (based on the mean distance determined by the data mining algorithm in Jabbar et al., 2018). We reclassified the kernel density maps into binaries (1 = exposed [any amount], 0 = not exposed) and multiplied the number of chemicals as informed by the top association patterns. Then, we assigned the pattern overlays to the maternal postal codes (Fig. 3). Additional data from the years 2013–2014 were used in the same manner for post-hoc independent assessment.

2.13. Assessment of the statistical support of the secondary patterns

From the previous step, we classified each birth having an ABO or non-ABO (1 = yes, 0 = no) and whether the mother was exposed or not to the mixture (1 = yes, 0 = no). Given the binary nature of the outcome, we applied multiple logistic regression models to evaluate the statistical significance of the pattern *mixture → ABO*. We conducted regression models specified for each ABO and the chemical mixture identified in the previous step after adjusting by specific ABO-related maternal factors, NO_2 concentrations, area-level SES, and main effects of single chemicals (as binary variables). We used previously selected maternal variables associated with each ABO (Serrano-Lomelin, 2017) as listed in Tables 1 and 2. Given the exploratory nature of the study, we exclusively interpreted the statistical significance of the ORs of the patterns as the means to add statistical support to hypotheses, to inspire future research. Logistic models were done using Stata 12 (StataCorp, 2011).

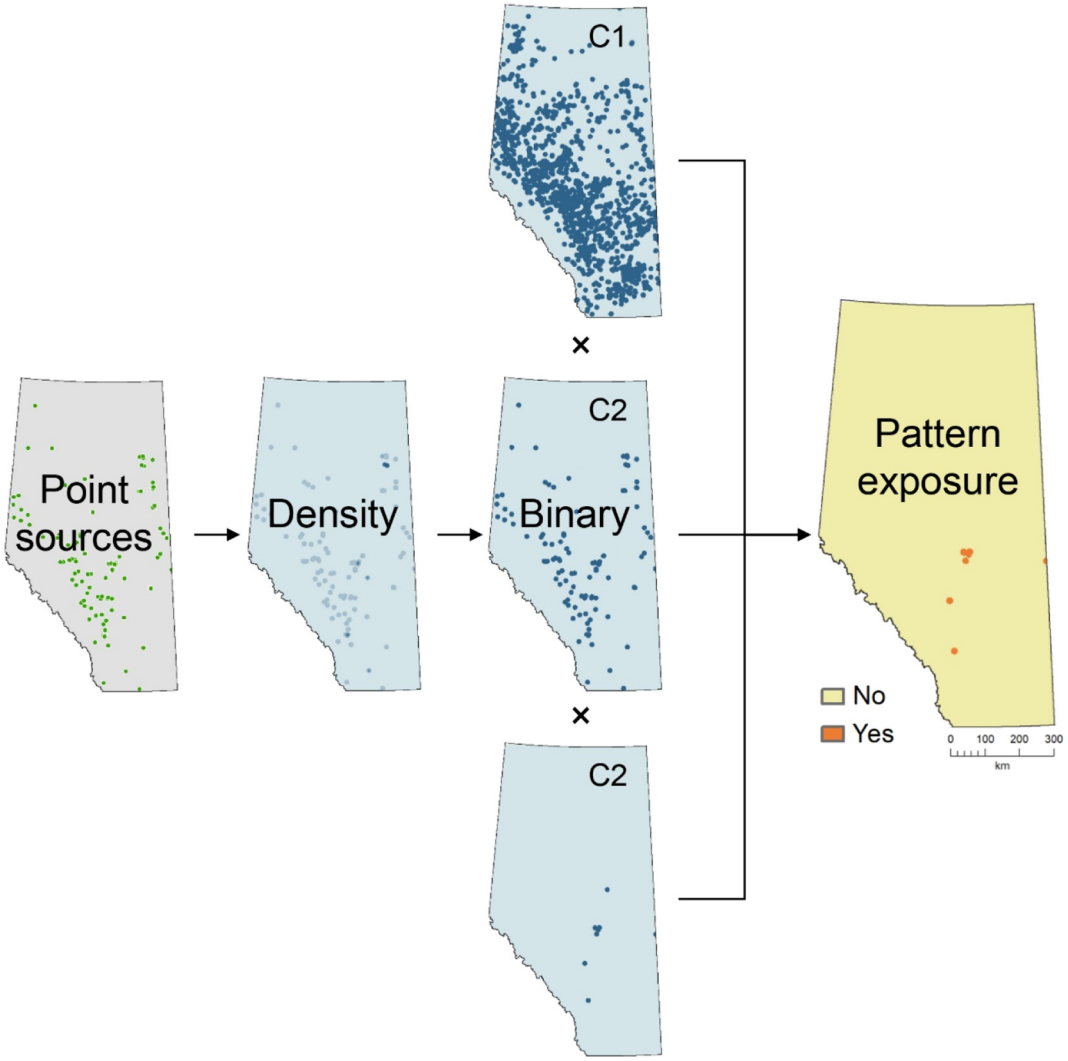

Fig. 3. The geographic information systems calculated each chemical emission as a density in tonnes/km^2 reclassified to binary, and then multiplied the three chemicals (C1, C2, C3) into the combined mixture patterns and assigned the potential exposure to the postal code of the maternal residences.

Table 1
Odds Ratios for the industrial air pollutant mixtures of three chemicals and preterm birth (PTB), small for gestational age (SGA), and low birth weight at term (LBWT), as selected from spatial data mining patterns according to lift ratio. 2006–2012.

Patterns	PTB				SGA				LBWT			
Chemical	Lift PTB	Lift non-PTB	LR	Adj-OR[a] [95%CI]	Lift SGA	Lift non-SGA	LR	Adj-OR[b] [95%CI]	Lift LBWT	Lift non-LBWT	LR	Adj-OR[b] [95%CI]
Mixture 1	9.02	7.48	1.21	1.20 [1.14, 1.27]	8.83	7.41	1.19	1.27 [1.21, 1.33]	12.09	7.44	1.63	1.26 [1.12, 1.40]
PM[c]	1.22	1.21	1.01	1.08 [1.03, 1.14]	1.22	1.21	1.01	1.22 [1.17, 1.28]	1.35	1.21	1.12	1.19 [1.07, 1.32]
Methyl ethyl ketone	7.77	6.46	1.2	1.17 [1.13, 1.20]	7.61	6.38	1.19	1.07 [1.04, 1.10]	10.15	6.42	1.58	1.08 [1.02, 1.15]
Xylene	2.22	2.16	1.03	1.08 [1.05, 1.12]	2.22	2.15	1.03	1.10 [1.08, 1.13]	2.77	2.15	1.29	1.16 [1.10, 1.23]
Mixture 2	9.11	7.55	1.21	1.20 [1.13, 1.26]					12.36	7.51	1.65	1.25 [1.12, 1.40]
PM[c]	1.22	1.21	1.01	1.08 [1.03, 1.14]					1.35	1.21	1.12	1.19 [1.07, 1.32]
Methyl ethyl ketone	7.77	6.46	1.2	1.17 [1.13, 1.20]					10.15	6.42	1.58	1.08 [1.02, 1.15]
Toluene	1.86	1.78	1.04	1.14 [1.11, 1.18]					2.34	1.77	1.32	1.14 [1.08, 1.21]
Mixture 3	8.55	7.04	1.22	1.24 [1.13, 1.35]	8.6	7.02	1.22	1.23 [1.13, 1.33]	11.38	7	1.63	1.28 [1.06, 1.53]
CO[d]												
PM[c]	1.22	1.21	1.01	1.08 [1.03, 1.14]	1.22	1.21	1.01	1.22 [1.17, 1.28]	1.35	1.21	1.12	1.19 [1.07, 1.32]
2-Butoxyethanol	8.38	6.92	1.21	1.14 [1.11, 1.18]	8.43	6.91	1.22	1.08 [1.05, 1.11]	10.99	6.89	1.6	1.09 [1.03, 1.16]
Mixture 4	8.62	7.11	1.21	1.15 [1.08, 1.23]	8.66	7.09	1.22	1.42 [1.34, 1.50]				
PM[c]	1.22	1.21	1.01	1.08 [1.03, 1.14]	1.22	1.21	1.01	1.22 [1.17, 1.28]				
Xylene	2.22	2.16	1.03	1.08 [1.05, 1.12]	2.22	2.15	1.03	1.10 [1.08, 1.13]				
n-Butyl alcohol	8.56	7.05	1.21	1.07 [1.02, 1.12]	8.59	7.04	1.22	1.19 [1.14, 1.23]				
Mixture 5	8.69	7.14	1.22	1.19 [1.08, 1.30]	8.72	7.12	1.22	1.36 [1.25, 1.48]				
PM[c]	1.22	1.21	1.01	1.08 [1.03, 1.14]	1.22	1.21	1.01	1.22 [1.17, 1.28]				
CO[d]												
n-Butyl alcohol	8.56	7.05	1.21	1.07 [1.02, 1.12]	8.59	7.04	1.22	1.19 [1.14, 1.23]				

Empty table cells indicate that the mixture did not spatially co-locate with that particular ABO.

[a] Odds ratio adjusted for maternal age, past-preterm, bleeding anytime, gestational hypertension, gestational diabetes, smoking during pregnancy, substance use during pregnancy, SES-index, multiparity, NO_2, and main effects of single chemicals.

[b] Odds ratio adjusted for maternal age, past-SGA, gestational hypertension, pre-pregnancy maternal weight < 45 kg, smoking during pregnancy, substance use during pregnancy, SES-index, multiparity, NO_2.

[c] PM = $PM_{2.5}$ or PM_{10} since the odds ratios were the same for both; therefore, the table has been simplified as PM to indicate either type.

[d] CO alone did not spatially co-locate with ABO cases.

Table 2
Odds Ratios for the industrial air pollutant mixtures of three chemicals and preterm birth (PTB), small for gestational age (SGA), and low birth weight at term (LBWT), using data from 2013 to 2014.

Patterns	PTB	SGA	LBWT
Chemicals	Adj-OR[a] [95%CI]	Adj-OR[b] [95%CI]	Adj-OR[b] [95%CI]
Mixture 1 (PM[c], methyl ethyl ketone, xylene)	1.04 [0.93, 1.17]	1.09 [0.99, 1.20]	**1.32 [1.06, 1.64]**
Mixture 2 (PM[c], methyl ethyl ketone, toluene)	1.11 [0.98, 1.24]		**1.37 [1.10, 1.71]**
Mixture 3 (CO, PM[c], 2-butoxyethanol)	1.00 [0.87, 1.16]	**1.26 [1.12, 1.42]**	**1.40 [1.06, 1.85]**
Mixture 4 (PM[c], xylene, n-butyl alcohol)	1.00 [0.89, 1.14]	**1.24 [1.12, 1.37]**	
Mixture 5 (PM[c], CO, n-butyl alcohol)	1.05 [0.91, 1.20]	**1.13 [1.16, 1.45]**	

Empty table cells indicate that the mixture did not spatially co-locate with that particular ABO.

[a] Odds ratio adjusted for maternal age, past-preterm, bleeding anytime, gestational hypertension, gestational diabetes, smoking during pregnancy, substance use during pregnancy, SES-index, multiparity, NO_2, and main effects of single chemicals.

[b] Odds ratio adjusted for maternal age, past-SGA, gestational hypertension, pre-pregnancy maternal weight < 45 kg, smoking during pregnancy, substance use during pregnancy, SES-index, multiparity, NO_2, and main effects of single chemicals.

[c] PM = $PM_{2.5}$ or PM_{10}.

3. Results

3.1. Descriptive statistics

The total number of births registered in Alberta from 2006 to 2012 was 349,762. Ninety-six percent (n = 336,588) were singleton live births with a gestational age between 22 and 42 completed weeks. Around 1% of these births (n = 3341) had an erroneous postal code and were excluded from the analysis. Therefore, we included a total of 333,247 singleton live births for spatial data mining analysis (Fig. 4). The total number of births from 2013 to 2014 was 108,547, and after applying the same criteria mentioned above, 103,551 were included in the subsequent assessment.

We extracted from the National Pollutant Release Inventory 62,641 entries of pollutants released to air in Alberta between 2006 and 2012. They included over six thousand facilities (n = 6279), geographically dispersed across the province (Fig. 5) that reported more than seven million tonnes of 136 different chemicals released into the air. The reported emissions of SO_2, NO_2, CO, and PM (either PM_{10} or $PM_{2.5}$) accounted for ≈97% of the total mass (tonnes) of air pollutant emissions, and the remaining 3% was composed of inorganics (≈1.4%), volatile organic compounds (VOCs) (≈1.3%), other organics (≈0.01%), metals (≈0.009%), nitrosamines/ethers/alcohols (≈0.004%), and polycyclic aromatic hydrocarbons (PAHs) (≈0.003%). For the subsequent assessment of the years 2013 to 2014, we extracted 10,615 entries of the seven pollutants identified in our top five secondary pattern (*mixtures*) emitted by 3249 facilities.

3.2. Primary and secondary co-location patterns

The AGT-Fisher algorithm discovered 1700 primary patterns (including *chemical → ABO* and *mixture → ABO*) from which we identified

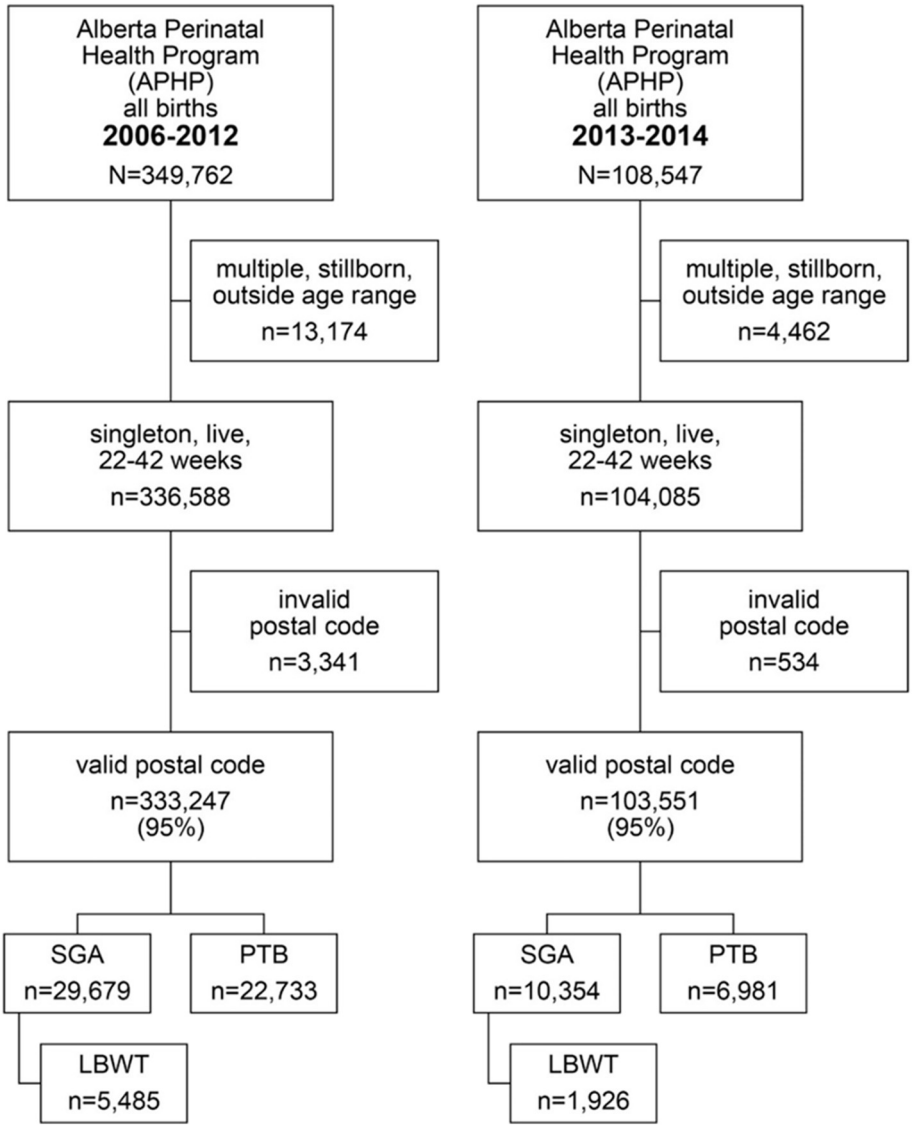

Fig. 4. The health outcome data were provided by the Alberta Perinatal Health Program, and the birth records were classified according to adverse birth outcome (ABO). The 2006–2012 data were used in the spatial data mining, and the 2013–2014 data were used to independently assess the patterns generated from the first set. ABO prevalence in the 2006–2012 and 2013–2014 periods was, respectively, as follows: PTB, 6.8% and 6.7%; SGA, 8.9% and 10%; LBWT, 1.7% and 1.9%.

the *mixture → ABO* patterns with the highest lift. A set of five secondary patterns with the highest LR criteria were selected for further analysis (Table 1). Their LRs were around 1.2 for PTB and SGA, and 1.6 for LBWT. One pattern did not apply to SGA, and two other patterns did not apply to LBWT and therefore were not included in the further statistical analysis assessment for those adverse birth outcomes (see Section 3.4).

3.3. Spatial exposure-areas to the five secondary patterns

Maps were instrumental for both visualization and analyses. Supplemental Figs. S1 and S2 present the following maps and descriptions: (i) contextual maps, including the SES-index and NO_2-land use regression model; and (ii) kernel densities of the chemicals identified in the five patterns. Supplemental Fig. S3 shows the location of the five pattern overlays used for assigning exposure to the maternal postal codes.

3.4. Statistical support to secondary patterns

Table 1 shows the lifts, LR, and adjusted ORs (with 95% CI) for the chemicals and mixtures of the five top lift mixtures. In all cases, the LR were higher than the lift of the single chemicals participating in these mixtures. The selected patterns were statistically significant after adjusting for the main effects of the single chemicals and other relevant covariates. The general term PM is used to simplify the presentation of the results by indicating either PM_{10} or $PM_{2.5}$ since the spatial co-location (exposure), and ORs were identical for the patterns including PM (i.e., the OR for PM_{10}/methyl ethyl ketone/xylene = $PM_{2.5}$/methyl ethyl ketone/xylene).

Overall, all *mixtures* showed significant statistical association with at least one ABO. We summarize the results for each *mixture* indicating the adjusted OR (adj-OR) and the number of cases (n) per ABO that were exposed to the mixture. *Mixture 1* (PM + methyl ethyl ketone + xylene) was positively associated with PTB (adj-OR = 1.20, n = 8102), SGA (adj-OR = 1.27, n = 10,466), and LBWT (adj-OR = 1.26, n = 1966). *Mixture 2* (PM + methyl ethyl ketone + toluene) was positively associated with PTB (adj-OR = 1.20, n = 8297) and LBWT (adj-OR = 1.25, n = 1993). *Mixture* 3 (CO + *PM* + 2-butoxyethanol) was positively associated with PTB (adj-OR = 1.24; n = 6950), SGA (adj-OR = 1.23, n = 9163), and LBWT (adj-OR = 1.28, n = 1717). *Mixture* 4 (PM + xylene + n-butyl alcohol) was positively associated with PTB (adj-OR = 1.15, n = 2499) and SGA

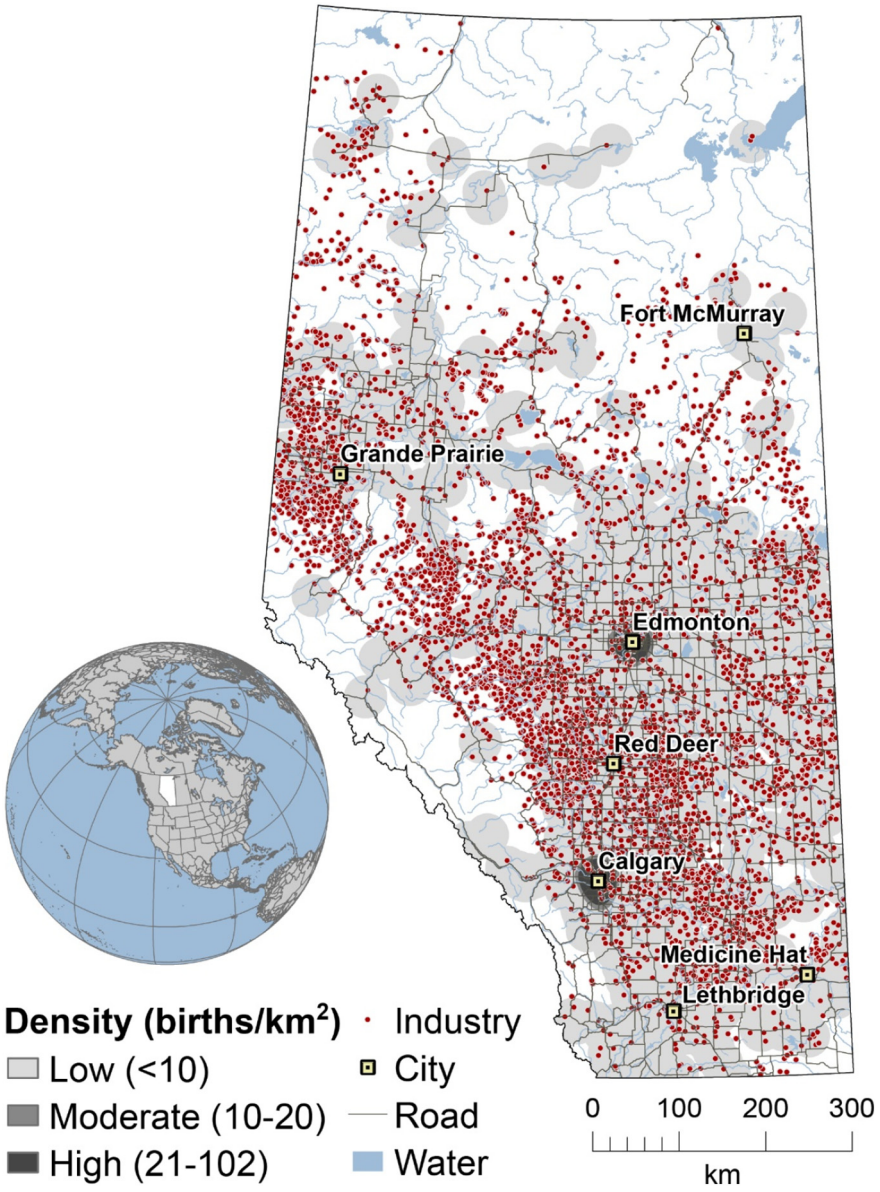

Fig. 5. The Alberta study area showing the distribution of the industrial facilities (red dots) and density of births, 2006–2012. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(adj-OR = 1.42, n = 3677). *Mixture* 5 (PM + CO + n-butyl alcohol) was positively associated with PTB (adj-OR = 1.19, n = 2485) and SGA (adj-OR = 1.36, n = 3665).

We subsequently assessed the odds ratios for mixtures 1 to 5 using data from 2013 to 2014, yielding similar results for SGA and LBWT (Table 2). For PTB, none of the mixtures were associated with it in 2013–2014.

4. Discussion

Assessing the impact of exposures to chemical mixtures of pollutants on human health represents a challenge (Feron et al., 2002). The scarcity of measured data and of appropriate methodological approaches remain as two of the principal obstacles. The availability of pollutant release and transfer registries or chemical biomonitoring programs has opened the door to simultaneously analyze hundreds of chemicals using data mining as a step forward (Bell and Edwards, 2015; Bellinger et al., 2017). In this study, we used a novel approach to generate new research hypotheses focusing on potential associations between ABO and hazardous mixtures of chemicals emitted by industrial sources. Our research findings were enabled by the iterative participation and contribution of researchers from different disciplines and knowledge users, and the integration of three different methodologies (data mining, GIS, and regression analysis).

4.1. Summarizing the process

The spatial data mining algorithm provided an initial and efficient search of the patterns from an enormous data space based on co-location of chemical mixtures and ABO (Jabbar et al., 2018). Since our primary interest was to identify *mixtures*, we focused on searching patterns with more than two chemicals. The primary patterns involving three chemicals, were those with the highest lift. Patterns, including four or more chemicals were rare (data not shown). We used the LR as the pruning criterion for selecting the secondary patterns assuming that any pregnant woman in a population is exposed to the *mixture*, but the lift was more likely to be higher when there was an adverse outcome. Finally, exposure assignment using geographic methods and statistical

evaluation of the patterns gave additional support to the candidate hypotheses.

4.2. The spatial data mining algorithm

Traditional data mining algorithms are designed to find associations in transaction datasets (Agrawal and Srikant, 1994). However, they do not have a temporal or geospatial component. Our algorithm was suitable to discover co-location patterns of spatial features, such as the mother's location and the existence of specific airborne chemicals, whose instances were often located together in spatial proximity. The robustness of our algorithm relies on the use of transactionalization in conjunction with the Kingfisher algorithm and Fisher's exact test. Transactionalization enables the use of robust methods such as the Kingfisher search method (a method that uses enumeration trees to find significant rules) in combination with the Fisher's exact test; the latter ensures that the antecedents (mixtures) and consequents (ABO) are statistically dependent (and not resulting by chance). The combination of these three methods reduced the number of spurious associations, even when a conventional p-value of 0.05 is used for extracting the initial patterns. However, establishing a p-value in data mining to find only theoretically-sound patterns is particularly difficult in the exploratory analysis focused on finding "new discoveries," which was the main objective of this research. For this reason, we introduced the use of the lift ratio as a theoretically-driven post-pruning process adding meaningful context to the patterns (secondary patterns). A complete discussion of the robustness (and other issues: e.g., the use of a 1 km grid points) of our spatial data mining approach have been recently published (Jabbar et al., 2018). Moreover, we went beyond the data mining approach by testing the significance of some of the secondary patterns using GIS (to assign exposure) and logistic models (which incorporated relevant risk factors to ABO: maternal variables and the SES-index) in two independent datasets (2006–2012 and 2013–2014).

4.3. From transactions to individuals

Once we selected the candidate hypotheses, the use of kernel densities identified the location (the maternal postal codes) where it was more probable that exposure during pregnancy occurred, allowing us to remove the obstacle of working with transactions. Kernel density has been proven as a useful method for estimating source-based exposures. It accounts for emission amounts and distance decay weighting for all point sources to spread the exposure in the 2-dimensional space (Jerrett et al., 2005; Wu et al., 2011; Stieb et al., 2016; Nielsen et al., 2017). Essentially, the tonnes emitted (i.e., because concentrations were not reported) within 10 km shaped the kernels, and the overlay of the three chemicals defined the exposure patterns. The exposure areas delineated by kernel densities, and assuming a mobility radius of 10 km of women during pregnancy, are prone to misclassification of exposure, an issue that is commented in the section of limitations.

The use of the LR was introduced, as we have mentioned, to add meaningful context to the patterns. LR is an objective measure that uses the familiar concept of ratio to compare the association *mixture* → ABO cases in relation to the association *mixture* → non-ABO cases. Thus, a LR value above 1, points to an association pattern and may serve as an effective measure for extracting good candidate hypotheses in transactional datasets. In fact, the lift ratio was found to be relevant for team members by its similarities with the odds ratio (OR) criteria, commonly used in environmental epidemiology. However, by using LR, it is possible to miss potentially relevant mixtures that co-locate only with cases (*mixture* → ABO cases) and not with non-cases. Thus, it is important to recognize that using other measures to objectively select association patterns may result in different top patterns. Some researchers using non-spatial data mining algorithms have estimated OR directly from the data mining algorithm (Toti et al., 2016), whereas others have selected patterns combining high values of *lift* and significant OR for pruning patterns (Park et al., 2014). Recently, Vu et al. (2019) undertook an empirical and theoretical examination of the relations between lift and odds ratio, founding a positive correlation. In our case, direct estimation of OR was not possible because the spatial data mining algorithm worked with transactions rather than cases. Thus, while the transactionalization improved the spatial approach, it limited the estimation of other relevant metrics such as the odds ratio (Jabbar et al., 2018).

4.4. The significance of the mixtures

The allocation of births to areas of exposure allowed testing associations using logistic regression. Interestingly, all of the five mixture patterns identified included particulate matter (PM). Of all routinely monitored ambient air pollutants, PM is one of the most commonly linked to PTB, SGA, and LBW (Malley et al., 2017; Lamichhane et al., 2015; Shah and Balkhair, 2011; Stieb et al., 2015). Our results extended these findings to industrial sources of PM. Two-mixture patterns included carbon monoxide, which had also previously been identified as associated with ABO in studies using monitored data (Stieb et al., 2012; Qian et al., 2016). The volatile organic compounds (VOCs), toluene, and xylene, included in the mixture patterns, have previously exhibited associations with ABO as members of the benzene, toluene, ethylbenzene, xylene (BTEX) group (Aguilera et al., 2010; Ghosh et al., 2012). Volatile organic compounds, as a group, have also been associated with ABO (Chang et al., 2017), but we are not aware of direct evidence explicitly linking the other three VOCs found in our mixture patterns, 2-butoxyethanol, n-butyl alcohol and methyl ethyl ketone with ABO. Interestingly, all, but n-butyl alcohol, represent recognized or suspected developmental toxicants (Office of Environmental Health Hazard Assessment: Resolution 65, 2018). When exploring for existing toxicological evidence on mixtures involving the seven chemicals of interest, Kim et al. reported toxicity potentiation when methyl ethyl ketone and toluene were combined (Kim et al., 2014). However, to our knowledge, this is the first time these chemicals were identified as part of patterns or mixtures in relation to ABO. The evidence cited here substantiates the value of the five hypotheses associating chemical mixtures with ABO described in this paper. Our subsequent assessment using data from another period also supports, except for PTB, the reproducibility of the findings. Differences in the amounts emitted for those chemicals between the two periods may account for the latter results (Supplemental Table S3). We corroborated that they were emitted and reported for the 2013–2014 period and that there were no changes in the NPRI's reporting requirements between the two periods (Environment and Climate Change Canada, 2019).

Our results support the idea that studying the health effects of mixtures may further our understanding of the relationships between air pollution and health. First, we observed that some chemicals were ubiquitous in the air and were often accompanied by other co-pollutants. For example, the spatial association of ABO and non-ABO with CO was statistically significant only when other chemicals accompanied CO. This observation suggests that the existing evidence linking individual monitored pollutants (e.g., PM, CO) with ABO, may be acting as proxies of more complex chemical mixtures, showing just the tip of the iceberg. Second, the *mixtures* could be conceptualized as entities with entirely different toxicity from the one derived by just adding the toxicity of the participating individual chemicals in the mix. Thus, the products of the reaction could represent a different toxic entity. A recent comprehensive review of the effects of ambient air pollution on pregnancy outcomes has pointed out inconsistencies in the effects described in related epidemiological studies (Klepac et al., 2018). These discrepancies could be related to variations in the toxicity derived from the conditions of the local mixtures of air pollutants. It has been shown, experimentally, that the interplay between PM-composition and PM-concentration results in a non-linear relation with PM-induced biological outcomes (Manzano-León et al., 2016). Finally, we want to emphasize that the significance of the chemical interactions within the

mixtures were not further analyzed as it would have required a different approach beyond the scope of this paper. The statistical analysis of the association rules tested here only provides statistical support as a way to advance them as robust hypotheses.

4.5. Limitations

Limitations of this study are mainly related to the assumptions and simplifications we made during the research process and to model bias.

Assumptions and simplifications were made based on the available data. The use of yearly estimations of chemical emissions limited our capacity to assign individual exposure to the mixtures better. Therefore, some misclassification bias in exposure is expected. The data on industrial air emissions from the National Pollutant Release Inventory are annual estimates and not monitored releases, which forces us to assume that they were equally present throughout the pregnancy period and years. This source of data is, however, the only industrial emission data available for a large number of chemicals emitted to the air. Other studies have shown the potential of using these data in health research (Wine et al., 2014). Besides, working with averaged annual exposures precludes the exploration of time-related windows of susceptibility during pregnancy, an important issue that should be taken into account, when possible, in perinatal and air-pollution epidemiology (Wang et al., 2018). Concerning this point, the use of spatial-temporal data mining algorithms may provide better insights. The algorithm we used can incorporate the temporal dimension when temporal data are available (e.g., using data from air monitoring stations), which is one of our targets for future research.

Another limitation is the spatial inaccuracy expected in rural areas: although we had access to all births for the study, we were limited to using the six-character postal codes, which tend to be vast areas in rural Alberta.

4.6. Model bias

We acknowledge that the discrepancy between spatial data mining and GIS methods when assigning exposures to the mothers may produce biased results. The 10 km circular buffer used in the GIS approximation only captured the area but not the directionality included in the spatial data mining. As a result, the estimated exposure areas, from both methods, did not match exactly. It is known that the propagation of the uncertainty from using different methods with different conceptualizations and model parameters may introduce bias in the results (Tasdighi et al., 2018). In order to reduce this bias, we used an independent data set (2013–2014) to tests the significance of the *mixtures* derived from the original data set (2006–2012). All *mixtures* showed significance with at least one ABO.

For all of the above, we want to emphasize that this study cannot imply causation, but rather provides a framework to effectively search through all combinations of chemicals to discover statistically supported patterns for further assessment. The spatial data mining algorithm can be applied to any health outcome that has potential spatial relationships with exposures.

5. Conclusions

We identified potentially hazardous mixtures of industrial pollutants spatially related to the occurrence of ABO in Alberta. The collaborative integration facilitated three outputs: (i) tools – the spatial data mining algorithm, filtering process to reduce all patterns to a more manageable number of patterns, and GIS methods to assign exposures; (ii) patterns – the associations of chemical mixtures and ABO; and (iii) hypotheses – selected subset of patterns for evaluation with conventional epidemiological methods. We extracted interesting association patterns (candidate hypotheses) from the spatial data mining, then performed geographic and statistical analyses to assign a measure of robustness to the selected hypotheses to further validate them as suitable research hypotheses. Data science is an exploratory process in which the significance of the discoveries improves when users interact with and explore the discovered patterns. Thus, we were able to test the plausibility of hypotheses derived from spatial data mining on the associations between industrial air pollutant mixtures and ABO, using an interdisciplinary methodological framework.

Pregnant mothers may be at higher risk for ABO when exposed to industrial chemical emissions in mixtures of PM, CO, xylene, toluene, methyl ethyl ketone, 2-butoxyethanol and n-butyl alcohol.

Future research in basic science (i.e., toxicological studies using cell-models or animal models) and observational studies using biomonitoring -to corroborate the presence of those chemicals in the body- is encouraged to test these hypotheses. Those studies can complement our understanding of the relationships between chemical mixtures and adverse birth outcomes.

Contributors

JSL, CN, SJ, OW, and CB performed the data management, analyses, and led the writing of the manuscript. OZ and AOV were the principal investigators of the project; all other team members provided interdisciplinary intellectual insights throughout project conception to end, interpretation, and reviewed and approved the final manuscript.

Declaration of Competing Interest

The authors declare they have no actual or potential competing financial interests. Funding agencies were not involved in any process of the research (study design; collection, analysis and interpretation of data; writing the report; decision for publication).

Acknowledgments

Funding: This work was supported by the Canadian Institutes of Health Research (CIHR) and The Natural Sciences and Engineering Research Council of Canada (NSERC), Funding Reference Number (FRN) 127789, for the Data Mining and Neonatal Outcomes (DoMiNO) Project, 2013–2018. Anonymized health data provided by the Alberta Perinatal Health Program (APHP). JSL was partially supported by CONACYT (Mexico). OW was partially supported by Women and Children Health Research Institute (University of Alberta).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2019.104972.

References

ACIS, 2010. AgroClimatic Information System AIS live Alberta weather station data. 2006–2010 [digital data]. Available: http://www.agric.gov.ab.ca/app116/stationview.jsp.

Agarwal, N., Banternghansa, C., Bui, L.T.M., 2010. Toxic exposure in America: estimating fetal and infant health outcomes from 14 years of TRI reporting. J. Health Econ. 29, 557–574. https://doi.org/10.1016/j.jhealeco.2010.04.002.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB. Vol. 1215. http://www.vldb.org/conf/1994/P487.PDF.

Aguilera, I., Garcia-Esteban, R., Iñiguez, C., Nieuwenhuijsen, M.J., Rodríguez, À., Paez, M., et al., 2010. Prenatal exposure to traffic-related air pollution and ultrasound measures of fetal growth in the INMA Sabadell cohort. Environ. Health Perspect. 118, 705–711. https://doi.org/10.1289/ehp.0901228.

Alberta Health Services, 2014. Alberta Perinatal Health Program, 2006–2014 [digital data]. http://aphp.dapasoft.com.

Auger, N., Authier, M.-A., Martinez, J., Daniel, M., 2009. The association between rural-urban continuum, maternal education and adverse birth outcomes in Québec, Canada. J. Rural. Health 25, 342–351. https://doi.org/10.1111/j.1748-0361.2009.00242.x.

Bell, S.M., Edwards, S.W., 2015. Identification and prioritization of relationships between

environmental stressors and adverse human health impacts. Environ. Health Perspect. 123, 1193–1199. https://doi.org/10.1289/ehp.1409138.
Bellinger, C., Jabbar, M.S.M., Zaïane, O.R., Osornio-Vargas, A., 2017. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health 17, 1–19. https://doi.org/10.1186/s12889-017-4914-3.
Casey, J.A., Savitz, D.A., Rasmussen, S.G., Ogburn, E.L., Pollak, J., Mercer, D.G., et al., 2016. Unconventional natural gas development and birth outcomes in Pennsylvania, USA. Epidemiology. 27, 163–172. https://doi.org/10.1097/EDE.0000000000000387.
Chan, E., Serrano, J., Chen, L., Stieb, D.M., Jerrett, M., Osornio-Vargas, A., 2015. Development of a Canadian socioeconomic status index for the study of health outcomes related to environmental pollution. BMC Public Health 15, 714. https://doi.org/10.1186/s12889-015-1992-y.
Chang, M., Park, H., Ha, M., Hong, Y.C., Lim, Y.H., Kim, Y., et al., 2017. The effect of prenatal TVOC exposure on birth and infantile weight: the mothers and children's environmental health study. Pediatr. Res. 82, 423–428. https://doi.org/10.1038/pr.2017.55.
CIHR. Canadian Institutes of Health Research. 2012. Guide to knowledge translation planning at CIHR: Integrated and End-of-Grant Approaches. Canada, ON. Available: http://www.cihr-irsc.gc.ca/e/documents/kt_lm_ktplan-en.pdf (Available at Jan 15, 2018).
Coker, E., Liverani, S., Ghosh, J.K., Jerrett, M., Beckerman, B., Li, A., et al., 2016. Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County. Environ. Int. 91, 1–13. https://doi.org/10.1016/j.envint.2016.02.011.
Currie, J., Schmieder, J., 2009. Fetal exposure to toxic releases and infant health. American Economic Association Papers and Proceedings (May) 177–183. https://doi.org/10.1257/aer.99.2.177.
DMTI Spatial, 2014. Platinum postal suite: CanMap multiple enhanced postal code geography 2001–2013. [digital data]. DMTI Spatial Inc., Markham, Ontariohttps://www.dmtispatial.com/canmap/.
Dominici, F., Peng, R.D., Barr, C.D., Bell, M.L., 2010. Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. Epidemiology. 21, 187–194. https://doi.org/10.1097/EDE.0b013e3181cc86e8.
Edwards, S., Maxson, P., Sandberg, N., Miranda, M.L., 2015. Air pollution and pregnancy outcomes. In: Nadadur, S.S., Hollingsworth, J.W. (Eds.), Air Pollution and Health Effects. Springer-Verlag, London, pp. 51–91 Chapter 3.
Environment and Climate Change Canada, 2015. The National Pollutant Release Inventory. Available. https://www.ec.gc.ca/inrp-npri/.
Environment and Climate Change Canada, 2016. National Pollutant Release Inventory. Summary report. In: Available, . http://publications.gc.ca/collections/collection_2017/eccc/En81-14-2016-eng.pdf.
Environment and Climate Change Canada, 2019. History of reporting requirements: National Pollutant Release Inventory. In: Available, . https://www.canada.ca/en/environment-climate-change/services/national-pollutant-release-inventory/substances-list/history-reporting-requirements.html.
Esri, 2016. ArcGIS Desktop, Release 10.5 [software]. Esri Inc.. https://www.esri.com.
Feron, V.J., Cassee, F.R., Groten, J.P., van Vliet, P.W., van Zorge, J.A., 2002. International issues on human health effects of exposure to chemical mixtures. Environ. Health Perspect. 110 (Suppl. 6), 893–899 doi:10.1289%2Fehp.02110s6893.
Ghosh, J.K.C., Wilhelm, M., Su, J., Goldberg, D., Cockburn, M., Jerrett, M., et al., 2012. Assessing the influence of traffic-related air pollution on risk of term low birth weight on the basis of land-use-based regression models and measures of air toxics. Am. J. Epidemiol. 175, 1262–1274. https://doi.org/10.1093/aje/kwr469.
Giudice, L.C., 2016. Environmental toxicants: hidden players on the reproductive stage. Fertil. Steril. 106, 791–794.
Hamalainen, W., 2012. Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. Knowl. Inf. Syst. 32, 383–414. https://doi.org/10.1007/s10115-011-0432-2.
Heaman, M., Kingston, D., Chalmers, B., Sauve, R., Lee, L., Young, D., 2013. Risk factors for preterm birth and small-for-gestational-age births among Canadian women: risk factors for PTB and SGA births. Paediatr. Perinat. Epidemiol. 27, 54–61 doi:10.1186%2Fs12884-016-0900-5.
Hidy, G.M., Pennell, W.T., 2010. Multipollutant air quality management. J Air Waste Manag Assoc. 60, 645–674. https://doi.org/10.3155/1047-3289.60.6.645.
Hystad, P., Setton, E., Cervantes, A., Poplawski, K., Deschenes, S., Brauer, M., et al., 2011. Creating national air pollution models for population exposure assessment in Canada. Environ. Health Perspect. 119, 1123–1129. https://doi.org/10.1289/ehp.1002976.
Jabbar, M.S.M., Bellinger, C., Zaiane, O.R., Osornio-Vargas, A., 2018. Discovering co-location patterns with aggregated spatial transactions and dependency rules. Int J Data Sci Anal. 5, 137–154. https://doi.org/10.1007/s41060-017-0079-5.
Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., et al., 2005. A review and evaluation of intraurban air pollution exposure models. J. Expo. Anal. Environ. Epidemiol. 15, 185–204. https://doi.org/10.1038/sj.jea.7500388.
Kapraun, D.F., Wambaugh, J.F., Ring, C.L., Tornero-Velez, R., Setzer, R.W., 2017. A method for identifying prevalent chemical combinations in the U.S. population. Environ. Health Perspect. 125, 087017. https://doi.org/10.1289/EHP1265.
Kim D, Saada A. 2013. The social determinants of infant mortality and birth outcomes in Western developed nations: a cross-country systematic review. 2013. Int J of Environ Res Public Health. 10, :2296–2335. doi:https://doi.org/10.3390/ijerph10062296.
Kim, K.W., Won, Y.L., Park, D.J., Kim, D.H., Song, K.Y., 2014. Comparative study on the EC50 value in single and mixtures of dimethylformamide, methyl ethyl ketone, and toluene. Toxicol Res. 30, 199–204. https://doi.org/10.5487/TR.2014.30.3.199.
Klepac, P., Locatelli, I., Korošec, S., Künzli, N., Kukec, A., 2018. Ambient air pollution and pregnancy outcomes: a comprehensive review and identification of environmental public health challenges. Environ Res J. 167, 144–159. https://doi.org/10.1016/j.envres.2018.07.008.
Kramer, M.S., 2003. The epidemiology of adverse pregnancy outcomes: an overview. J. Nutr. 133, 1592S–1596S. https://doi.org/10.1093/jn/133.5.1592S.
Kramer, M.S., Platt, R.W., Wen, S.W., Joseph, K.S., Allen, A., Abrahamowicz, M., et al., 2001. A new and improved population-based Canadian reference for birth weight for gestational age. Pediatrics 108, e35 PubMed ID: 11483845.
Lamichhane, D.K., Leem, J.-H., Lee, J.-Y., Kim, H.-C., 2015. A meta-analysis of exposure to particulate matter and adverse birth outcomes. Environ Health Toxicol., e2015011. https://doi.org/10.5620/eht.e2015011.
Li, J., Zaïane, O.R., Osornio-Vargas, A., 2014. Discovering statistically significant co-location rules in datasets with extended spatial objects. In: Data Warehousing and Knowledge Discovery. Springer, pp. 124–135.
Li, J., Adilmagambetov, A., Jabbar, M.S., Zaïane, O.R., Osornio-Vargas, A., Wine, O., 2016. On discovering co-location patterns in datasets: a case study of pollutants and child cancers. GeoInformatica. 20, 1–42. https://doi.org/10.1007/s10707-016-0254-1.
Malley, C.S., Kuylenstierna, J.C.I., Vallack, H.W., Henze, D.K., Blencowe, H., Ashmore, M.R., 2017. Preterm birth associated with maternal fine particulate matter exposure: a global, regional and national assessment. Environ Int. 101, 173–182. https://doi.org/10.5620/eht.e2015011.
Manzano-León, N., Serrano-Lomelin, J., Sánchez, B.N., Quintana-Belmares, R., Vega, E., Vázquez-López, I., Rojas-Bracho, L., et al., 2016. TNFα and IL-6 responses to particulate matter in vitro: variation according to PM size, season, and polycyclic aromatic hydrocarbon and soil content. Environ. Health Perspect. 124, 406–412.
Mauderly, J.L., Burnett, R.T., Castillejos, M., Özkaynak, H., Samet, J.M., Stieb, D.M., et al., 2010. Is the air pollution health research community prepared to support a multipollutant air quality management framework? Inhal. Toxicol. 22 (S1), 1–19. https://doi.org/10.3109/08958371003793846.
Nielsen, C.C., Amrhein, C.G., Osornio-Vargas, A.R., 2017. Mapping outdoor habitat and abnormally small newborns to develop an ambient health hazard index. Int. J. Health Geogr. 16, 1–21. https://doi.org/10.1186/s12942-017-0117-5.
Office of Environmental Health Hazard Assessment (OEHHA) Proposition 65. California: State of California Environmental Protection Agency. Available: https://oehha.ca.gov/proposition-65 (Available at July 11, 2018).
Park, S.H., Jang, S.Y., Kim, H., Lee, S.W., 2014. An association rule mining-based framework for understanding lifestyle risk behaviors. Tu Y-K. PLoS One 9, e88859. https://doi.org/10.1371/journal.pone.0088859.
Public Health Agency of Canada (PHAC). 2013. Perinatal Health Indicators for Canada 2013: A Report of the Canadian Perinatal Surveillance System. Ottawa. Available: http://publications.gc.ca/site/eng/411563/publication.html
Qian, Z., Liang, S., Yang, S., Trevathan, E., Huang, Z., Yang, R., et al., 2016. Ambient air pollution and preterm birth: a prospective birth cohort study in Wuhan, China. Int. J. Hyg. Environ. Health 219, 195–203. https://doi.org/10.1016/j.ijheh.2015.11.003.
SCHER (Scientific Committee on Health and Environmental Risks). Toxicity and Assessment of Chemical Mixtures. 2012. Available: https://ec.europa.eu/health/sites/health/files/scientific_committees/environmental_risks/docs/scher_o_155.pdf (Available at Jan 15, 2018). doi:https://doi.org/10.2772/21444
Serrano-Lomelin J. 2017. Profiling industrial air-pollutant mixtures and their associations with preterm birth and small for gestational age in Alberta, Canada. University of Alberta. 2017. Available: https://era.library.ualberta.ca/items/3fdd6a2e-454c-406f-a531-cecf487d3700/view/f565dcfb-d046-4f6a-8a9d-af114c660496/Serrano_Jesus_A_201712_PhD.pdf.
Shah, P.S., Balkhair, T., 2011. Air pollution and birth outcomes: a systematic review. Environ. Int. 37, 498–516. https://doi.org/10.1016/j.envint.2010.10.009.
Silverman, B.W., 1997. Density Estimation for Statistics and Data Analysis. Chapman and Hall.
Slama, R., Darrow, L., Parker, J., Woodruff, T.J., Strickland, M., Nieuwenhuijsen, M., et al., 2008. Meeting report: atmospheric pollution and human reproduction. Environ. Health Perspect. 116, 791–798. https://doi.org/10.1289/ehp.11074.
StataCorp. 2011. Stata Statistical Software: Release 12 [software]. College Station, TX: StataCorp LP. http://www.stata.com/
Statistics Canada. 2007. Census - Boundary files: Reference Guide. Statistics Canada Catalogue no. gda_000b06a_e.zip [accessed 5 Jun 2015]. Available: http://www12.statcan.ca/census-recensement/2011/geo/bound-limit/bound-limit-2006-eng.cfm
Statistics Canada. 2008. Census Dictionary. Statistics Canada Catalogue no. 92-566-XWE. Available: http://www12.statcan.ca/english/census06/reference/dictionary/index.cfm
Statistics Canada. 2012. Table 102-0562 - Leading Causes of Death, Infants, by Sex, Canada, Annual, 2006-2012 [Digital Data]. Canadian Socio-Economic Information Management System (CANSIM). Available: http://www5.statcan.gc.ca/cansim/pick-choisir?lang=eng&searchTypeByValue=1&id=1020562
Statistics Canada. 2014. Table 102-4318 - Birth-related Indicators (Low and High Birth Weight, Small and Large for Gestational Age, Pre-term births), by Sex, Three-year Average, Canada, Provinces, Territories, Census Metropolitan Areas and Metropolitan Influence Zones, Occasional [Digital Data]. Canadian Socio-Economic Information Management System (CANSIM). Available: http://www5.statcan.gc.ca/cansim/a05?lang=eng&id=01024318
Stieb, D.M., Chen, L., Eshoul, M., Judek, S., 2012. Ambient air pollution, birth weight and preterm birth: a systematic review and meta-analysis. Environ. Res. 117, 100–111. https://doi.org/10.1016/j.envres.2012.05.007.
Stieb, D.M., Chen, L., Beckerman, B.S., Jerrett, M., Crouse, D.L., Omariba, D.W., et al., 2015. Associations of pregnancy outcomes and PM in a national Canadian study. Environ. Health Perspect. 124, 243–249. https://doi.org/10.1289/ehp.1408995.
Stieb, D.M., Chen, L., Hystad, P., Beckerman, B.S., Jerrett, M., Tjepkema, M., et al., 2016. A national study of the association between traffic-related air pollution and adverse pregnancy outcomes in Canada, 1999–2008. Environ. Res. 148, 513–526. https://doi.

org/10.1016/j.envres.2016.04.025.
Sutton, P., Woodruff, T.J., Perron, J., Stotland, N., Conry, J.A., Miller, M.D., et al., 2012. Toxic environmental chemicals: the role of reproductive health professionals in preventing harmful exposures. Am. J. Obstet. Gynecol. 207, 164–173. https://doi.org/10.1016/j.ajog.2012.01.034.
Tasdighi, A., Arabi, M., Harmel, D., Line, D., 2018. A Bayesian total uncertainty analysis framework for assessment of management practices using watershed models. Environ Modell Soft. 108, 240–252. https://doi.org/10.1016/j.envsoft.2018.08.06.
Toti, G., Vilalta, R., Lindner, P., Lefer, B., Macias, C., Price, D., 2016. Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. Artif. Intell. Med. 74, 44–52. https://doi.org/10.1016/j.artmed.2016.11.003.
Tough, S.C., Svenson, L.W., Johnston, D.W., Schopflocher, D., 2001. Characteristics of preterm delivery and low birth weight among 113,994 infants in Alberta: 1994–1996. Can J Public Health. 92, 276–280. 11962113.
Vu, K., Clark, R.A., Bellinger, C., Erickson, G., Osornio-Vargas, A., Zaiane, O., Yuan, Y., 2019. The index lift in data mining has a close relationship with the association measure relative risk in epidemiological studies. BMC Med Inform Decis Mak. 19, 1–8. https://doi.org/10.1186/s12911-019-0838-4.
Walker Whitworth, K., Kaye Marshall, A., Symanski, E., 2018. Drilling and production activity related to unconventional gas development and severity of preterm birth. Environ. Health Perspect. 126, 037006. https://doi.org/10.1289/EHP2622.
Wang, A., Padula, A., Sirota, M., Woodruff, T.J., 2016. Environmental influences on reproductive health: the importance of chemical exposures. Fertil. Steril. 106, 905–929. https://doi.org/10.1016/j.fertnstert.2016.07.1076.
Wang, Q., Benmarhnia, T., Zhang, H., Knibbs, L.D., Sheridan, P., Li, C., et al., 2018. Identifying windows of susceptibility for maternal exposure to ambient air pollution and preterm birth. Environ. Int. 121, 317–324. https://doi.org/10.1016/j.envint.2018.09.021.
Wigle, D.T., Arbuckle, T.E., Turner, M.C., Bérubé, A., Yang, Q., Liu, S., et al., 2008. Epidemiologic evidence of relationships between reproductive and child health outcomes and environmental chemical contaminants. J Toxicol Environ Health B. 11 (5–6), 373–517. https://doi.org/10.1080/10937400801921320.
Wilhelm, M., Ghosh, J.K., Su, J., Cockburn, M., Jerrett, M., Ritz, B., 2012. Traffic-related air toxics and term low birth weight in Los Angeles County, California. Environ. Health Perspect. 120, 132–138. https://doi.org/10.1289/ehp.1103408.
Williams RG. 1999. Nonlinear surface interpolations: Which way is the wind blowing? In: Proceedings of 1999 ESRI International User Conference. http://proceedings.esri.com/library/userconf/proc99/proceed/papers/pap122/p122.htm
Willis M, Hystad P. 2019. Environmental Hazardous Air Pollutants and Adverse Birth Outcomes in Portland, OR. Environmental Epidemiology. 3, e034.:1–8; doi:https://doi.org/10.1097/EE9.0000000000000034.
Wine, O., Hackett, C., Campbell, S., et al., 2014. Using pollutant release and transfer register data in human health research: a scoping review. Environ. Rev. 22, 51–65. https://doi.org/10.1139/er-2013-0036.
Wine, O., Zaiane, O., Osornio Vargas, A.R., on behalf of the DoMiNO Project team, 2019. A collaborative research exploration of pollutant mixtures and adverse birth outcomes by using innovative spatial data mining methods: the DoMiNO project. Challenges. 10. In press, accepted(March 18, 2019).
World Health Organization, 2018. SDG 3: ensure healthy lives and promote wellbeing for all at all ages. Available: http://www.who.int/sdg/targets/en/.
Wu, J., Wilhelm, M., Chung, J., Ritz, B., 2011. Comparing exposure assessment methods for traffic-related air pollution in an adverse pregnancy outcome study. Environ. Res. 111, 685–692. https://doi.org/10.1016/j.envres.2011.03.008.