AI Applied to Air Pollution and Environmental Health: A Case Study on Hypothesis Generation

Colin Bellinger, Mohomed Shazan Mohomed Jabbar, Osnat Wine, Charlene Nielsen, Jesus Serrano-Lomelin, Alvaro Osornio-Vargas, and Osmar R. Zaïane

1 Introduction

Exposure to pollution in the environment is a major contributor to disease globally and has economic impacts on the order of billions of dollars each year [45]. Related to this, the field of environmental health aims to monitor and understand factors in the environment that affect human health and disease. This chapter examines the challenges related to understanding airborne chemical dispersion and human exposure along with the resulting adverse health outcomes and discusses how AI contributes to these tasks.

One of the first challenges in environmental health is understanding which airborne chemicals are present and at what levels. Physical pollution models are a standard method to estimate these quantities. Physical models are developed from domain expertise, along with data on the emission sources and chemicals of interest.

C. Bellinger (✉)
Digital Technologies, National Research Council of Canada, Ottawa, Canada
e-mail: colin.bellinger@nrc-cnrc.gc.ca

M. S. M. Jabbar · O. R. Zaïane
Alberta Machine Intelligence Institute, Edmonton, Canada
e-mail: mohomedj@ualberta.ca

O. R. Zaïane
e-mail: zaiane@ualberta.ca

O. Wine · A. Osornio-Vargas
Department of Pediatrics, University of Alberta, Edmonton, Canada
e-mail: osnat@ualberta.ca

A. Osornio-Vargas
e-mail: osornio@ualberta.ca

C. Nielsen
School of Public Health, University of Alberta, Edmonton, Canada
e-mail: ccn@ualberta.ca

J. Serrano-Lomelin
Department of Obstetrics and Gynecology, University of Alberta, Edmonton, Canada
e-mail: jaserran@ualberta.ca

F. Chen and J. Zhou (eds.), *Humanity Driven AI*,
https://doi.org/10.1007/978-3-030-72188-6_10

These are combined with atmospheric and meteorological factors that determine the transportation and evolution of the pollutants in the environment. Models of this nature are limited by the complexity in their design and prediction errors resulting from abstraction. Alternatively, geo-statistical approaches rely on multi-variate linear models that require large-scale spatial monitoring and geographic information systems (GIS) data to model and predict pollutant concentrations. The GIS parameters needed for model development and application are often available on a limited spatial scale, and models cannot generalise across cities [82]. In cases where the necessary data does exist, it is often held by separate institutions and corporations with distinct legal, moral and financial obligations. As a result, the datasets are often small, and the replications of studies are difficult. Nonetheless, there is a growing awareness in the environmental health community about the need for high-quality, accessible data [83]. This shift has opened the door for new and more power data-driven AI methods to play a role.

In addition to air pollution modelling and prediction, there is great need to advance the understanding of the health consequence of exposure to airborne pollutants. The chemicals in the atmosphere co-occur and exist as mixtures that interact with each other. Once inhaled, they persist in the body for varying lengths of time, which, amongst other things, depends on the chemical makeup. Recent evidence suggests that mixtures of chemicals can have a toxicological behaviour that differs from the toxicity of the individual chemicals [20, 77] and may produce greater adverse health outcomes [64, 71]. As a result, there is a growing movement in the environmental health community, including regulators, epidemiologists and health practitioners, to encourage the development of new paradigms of analysis to explore the impact of exposure to mixtures of airborne chemicals on health outcomes [24, 41, 57, 68]. The authors in [13, 57, 75] noted that the traditional tools of analysis are often insufficient to assess the impacts of mixtures of pollutants. There is a strong need for methods that can address the unique challenges presented by high-dimensional (multi-chemical) environmental health data [56]. In addition, there is a need for creative methods to fill the persistent data gaps related to movement and exposure, along with methods that can be applied in rural areas where data is even more sporadic.

Artificial intelligence (AI) is the computational process in which algorithms learning from data or experience, and are applied to analyse large datasets, discover patterns, extract actionable knowledge and predict outcomes of future or unknown events [5, 30]. Methods used in this process come from a combination of computational disciplines including statistics, mathematics, machine learning and database systems. Prior to the application of the AI algorithms, processing steps are often applied to format and clean the data. In addition, a post-processing stage is typically employed to visualise the results of the analysis in an intuitive and easy-to-communicate manner.

AI provides a wide array of scalable and reliable methods that have performed well in complex domains with similar challenges to those in environmental health. When paired with accurate data and domain expertise, AI algorithms have demonstrated a strong potential to support the advancement of knowledge and understanding in applications such as in science, engineering and medicine [35, 53, 69]. Moreover, new frontiers for the application of AI, which often inspire novel algorithms, analyses and evaluation methods, are being explored everyday. This has inspired collaborations between AI and environmental health researchers aimed at the adapting AI methods to analyse modern, big datasets in air pollution epidemiology [10, 61, 79, 82]. Researchers are now utilising the unique abilities of AI to incorporate new data sources, such as satellite and street view images, and social media posts into the analysis [67, 82]. The flexibility of AI has also been used to develop a better understanding of the impact of exposure to airborne chemical mixtures. A recent survey on machine learning applied environmental health found that 52% of the identified studies employed machine learning methods to analysing chemical mixtures [61].

The remaining of this chapter is laid out as follows. Section 2 presents four areas of environmental health related to air pollution in which AI has great potential. Specifically, air quality prediction and forecasting, health outcome analysis, source apportionment, and decision support. A case study of the use of AI to support the advancement of a particular application of environmental health is provided in Sect. 3. In particular, result from the Data Mining & Neonatal Outcomes (DoMiNO) project[1] is presented to illustrate how geo-spacial data mining and data visualisation can be combined with GIS and traditional epidemiological analysis to generate hypotheses about which mixtures of airborne chemicals negatively impact birth outcomes. Finally, Sect. 5 discusses implication, future work and challenges related to the use AI in environmental health related to air pollution and Sect. 6 summarises the finds of this chapter.

2 AI in Environmental Health

This section highlights four areas of environmental health related to air pollution in which AI has great potential. These include air quality prediction and forecasting, health outcome analysis, source apportionment and decision support.

2.1 Air Quality Prediction and Forecasting

A significant portion of the research on AI applied to environmental health deals with the challenge of forecasting and predicting airborne pollution levels. This includes predicting the current air quality or pollution levels, forecasting the future values, given some local or regional input variables, and forecasting the geo-spatial distribution of air quality or pollutants. Predictions of this nature serve to support public policy, planning and health research by simplifying and improving the accuracy of pollution estimates and contributing to the understanding of the impact of a potential future events, such as new highways and factories.

[1] Data MIning and Neonatal Outcomes: https://sites.google.com/a/ualberta.ca/domino/.

Some examples of the application of AI in air pollution prediction include [17, 46, 66, 86]. The authors in [66] compared traditional methods, such as generalised additive models, to the AI methods such as random forest (RF) and support vector machine (SVM) for predicting $PM_{2.5}$ during wildfire events. In [46], the authors evaluated the effectiveness of RF, SVM and artificial neural network (ANN) for estimating the daily distributions of $PM_{2.5}$. An ANN was employed in [17] to predict the indoor air quality based on data recorded at outdoor air monitors, and the authors in [49] employed boosted regression trees to improve the accuracy of common low-cost air pollution sensors.

$PM_{2.5}$ studies are typically limited to ground-based measurements. As a result, they often utilise land-use models to estimate the spatial distributions and exposure. Satellite-based data is expanding the spatial scope of the accessible data and enabling the incorporation of temporal analyses. Aerosol optical dept (AOD) data, collected as a part of NASA's earth observation program, has been used in combination with meteorological, atmospheric and land-use data to develop spatial-temporal $PM_{2.5}$ models [14]. In this work, RF models were trained to predict daily PM concentrations at a resolution of 1×1 km throughout the metropolitan area of Cincinnati, USA.

Whilst the majority of the previous work utilised traditional AI techniques from supervised and unsupervised learning, the power of deep learning is increasingly being recognised and exploited in state-of-the-art public health research [82]. Unlike physical and statistical models, classical methods from machine learning and deep learning methods have the potential to scale up to global coverage by exploiting the increasing supply of ground-based and satellite-based imagery, along with other remote sensing data. This is facilitated by deep learning's unique ability to efficiently generalise from large datasets composed of multiple data formats, such as image, text and sensor. Recently, researchers have utilised deep learning for haze prediction [52, 54] and for $PM_{2.5}$ and PM_{10} classification and exposure prediction [16, 22, 23]

The authors in [51] proposed the deep learning-based long short-term memory (LSTM) method to predict air pollutant concentrations at fixed locations based on historical air pollutant concentration data, meteorological data and other time series data. Their results suggest that the method can more effectively capture spatio-temporal correlations and incorporates auxiliary data to improve predictive performance. In addition to predicting outcomes, the proficiency of deep learning from image data provides the potential to identify prevalent co-occurring exposure "networks" through image recognition and unsupervised learning [62].

2.2 Health Outcome Analysis and Characterisation

A major challenge in environmental health is understanding the relationship between exposure to airborne chemicals and health outcomes. This challenge is exacerbated by the complexity of co-occurring airborne chemicals, the persistence of chemicals in the body for varying lengths of time and other risk factors.

In order to shed more light on the complex relationship between exposure and health outcomes, researchers are increasingly looking to existing and novel AI methods for help. The availability of pollutant release data, transfer registries and chemical biomonitoring data has opened the door to the application of AI to analyse large datasets of chemical exposures. The authors in [7], for example, used frequent itemset mining to efficiently and comprehensively evaluate relationships between chemicals and health biomarkers for diseases in the NHANES biomonitoring survey. In [39, 70], the authors developed a new co-location pattern mining algorithm AGT-Fisher (Aggregated Grid Transactionization) to discover spatial associations between mixtures of chemicals and adverse birth outcomes. In [77], an association rule mining-based methodology is used to discover patterns with relevant odds ratios whilst limiting redundancy and control for statistical significance. The author proposed a combined approach that first used AI to identify a subset of interesting associations between air pollutant exposure profiles and children's cognitive skills, and secondly, the approach utilised traditional statistical methods adjusted for confounders in order to estimate the magnitude. The two-stage approach is particularly effective for generating meaningful hypotheses within high-dimensional exposure data.

Unsupervised clustering algorithms are another effective method to help understand the relationship between airborne chemicals and health outcomes. The authors in [63] utilised self-organising maps (SOM) to identify pollutant profiles within the ambient air and associate them with health outcomes. This work improved the understanding of long-term spatial distributions of multiple pollutants and demographic characteristics of populations residing within areas with distinct air quality. Alternatively, K-means and hierarchical clustering were employed to group days with similar chemical profiles at a single site in Boston, USA [6]. The clusters described unique physical and chemical characteristics and are utilised to investigate physical and chemical conditions posing higher health risks. Bayesian clustering techniques are particularly interesting in the context of environmental health as they attempt to account for uncertainty in the data. The authors in [58] utilise Bayesian clustering to characterise the spatial distribution of multiple pollutants and populations at risk in Atlanta, USA.

2.3 Source Apportionment

Many countries regulate and require the reporting of chemicals emitted to the environment. Once released into the atmosphere, however, complex physical and chemical processes determine their fate. In addition, many chemical emissions, such as those from motor vehicles and aeroplanes, are not directly tracked. As a result, it is difficult to accurately associate local air quality measurements with the factors causing them. Source apportionment aims to trace a given decrease in air quality or increase in a given pollutant back to its emission source. Amongst other benefits, the ability to do so accurately can enable regulators to monitor emitters and is helpful for updating laws and taking steps towards mitigating the impact on humans and the broader environment.

Existing studies largely focus on outdoor and urban air pollution and apportioning particular airborne pollutants to potential sources, such as industrial sites, regions and major intersections. Clustering and data analysis methods have been applied to identify correlations and the importance of particular meteorological parameters, traffic, fuel fired equipment and industries to air pollution [18, 19, 73, 76, 80]. The authors in [87] utilised sequential pattern mining technique to investigate spatial-temporal patterns of $PM_{2.5}$.

2.4 Decision Support

As previously stated, accurate predictive models have the potential to support science and drive decision-making related to regulations and urban planning. Given its ability to incorporate multi-modal data, deep learning may serve as an efficient means of predicting past and future exposures based on known or anticipated changes in land use, traffic and the built environment. In addition, it may serve to identify areas to be prioritised for detailed monitoring and/or surveillance [82].

Existing work focused on discovering associations between chemical mixtures and health outcomes combined with source apportionment can serve to guide public policymakers to increase regulations on chemicals association with adverse health outcomes, work with neighbouring regions to reduce the impacts of upwind emissions and change industrial zoning to reduce the risk of the co-occurrence of chemicals that would form harmful mixtures. In addition, predictive models can be used to determine staffing and other public health needs. In [25], the authors use data from PM monitors to predict hospital admission for cardiovascular and respiratory diseases. Multiple data sources, including Twitter and Google searches, are utilised in [65] to predict asthma-related emergency department visits and can guide staffing levels.

3 Case Study: DoMiNO

This case study presents our interdisciplinary research with Data Mining & Neonatal Outcomes (DoMiNO) project.[2] This work serves to bridge the knowledge gap between exposure to airborne chemical mixtures during pregnancy and the occurrence adverse birth outcomes (ABOs). To achieve this, we utilise state-of-the-art methodologies from data mining and knowledge discovery. The developed spatial co-location pattern mining algorithm AGT-Fisher involves transforming the geo-

[2] Data MIning and Neonatal Outcomes: https://sites.google.com/a/ualberta.ca/domino/.

spatial pollution and birth outcome data into transactions for pattern mining with the Kingfisher algorithm. The Kingfisher algorithm discovers dependency rules of the form $A \rightarrow B$, where A is a set of airborne chemicals, and B is a birth outcome [38, 39, 50, 70].

In data-intensive applications, the data mining process often discovers a larger volume of patterns. The number of discovered patters is often greater than can be analysed and understood by the knowledge users. This poses a significant barrier to the effective utilisation of the mining results. In our work, for example, data mining with the AGT-Fisher algorithm [39] produced over 1700 statistically significant co-location patterns on our data with antecedents up to the size of three chemicals. Metrics of interestingness applied to sort the discovered patterns can only partially address this issue.

In order to facilitate the efficient use of the discovered patterns, we created the visualisation tool, Visualisation of Association Rules (VizAR). This tool advances upon the previous work by developing an interactive Web-based software platform for post-pattern mining, exploration and visualisation. Similar to the work of Ltifi et al. in [55], our goal is to support human intelligence with machine intelligence. Our work, however, focuses on geo-spatial environmental health data and the identification of valuable knowledge in mined co-location patterns.

VizAR serves as the final step in the data mining process, as illustrated in Fig. 1. The essential features of VizAR are (a) interactive exploration and (b) visualisations at multiple levels of geo-spatial abstraction. It operates on the mined patterns,end thereby alleviating the end-user from making complex technical decision regarding algorithms and metrics. It enables users to interactively search, sort, filter, explore and visualise the patterns and their geographic distribution at multiple levels of abstraction.

From a domain perspective, VizAR facilitates knowledge translation by enabling the users to connect the discovered pattern with its roots in the mined data. The results of this can both inspire new research questions and hypotheses and drive new public policy directions. In our results, we present two use-cases for the VizAR software that illustrate its ability to identify interesting and epidemiological significant patterns.

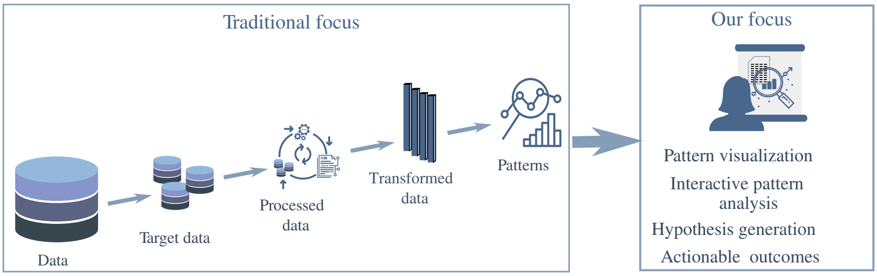


Fig. 1 This work focuses on the final step of the data science process. Specifically, the translation of patterns to actionable user knowledge. This image was inspired by one first appearing in [26]

We evaluate the meaningfulness of a subset of these patterns using the odds ratio, which is a standard approach in epidemiology.

Our evaluation demonstrates that the framework enables users to identify patterns that are pertinent to their work and chemical combinations for which the exposed group is at a greater risk than the unexposed according to the odds ratio. More generally, our user-base finds that pattern discovery via AGT-Fisher and presenting in VizAR enables them to identify new associations that have the potential to initiate research that could lead to healthier births in the future.

3.1 Related Work

In this work, we are interested in association rules, $A \rightarrow B$, where A is a set of airborne chemicals (i.e. antecedents), and B is an adverse birth outcome (i.e. consequent); in the geo-spatial context, these are referred to as co-location patterns. A co-location pattern is a set of spatial features whose instances are often located together in spatial proximity. Due to the significance in multiple fields of study, co-location pattern mining has gained significant importance recently [48]. We address the challenge of co-location pattern mining by transforming the geo-spatial data to a tabular format via aggregated grid transactionization. Transactionization enables the patterns to be discovered with standard association mining algorithms [4, 33].

Measures of interestingness play an essential role in the data mining process. These measures are intended for selecting and ranking patterns according to their potential interest to the user [40]. In addition, they are helpful for saving time and space costs associated with the data mining process [29]. Most of the existing association mining techniques rely on frequency-based prevalence or statistical significance to measure interestingness [4, 33, 85]. These include metrics such as support, confidence, lift and the p-value [29]. Because of the exploratory and interdisciplinary nature of data science, it is often challenging to select a metric that will accurately rank the patterns according to the users subjective preferences. Thus, to avoid pruning rules that may be of interest to the users, a low selection threshold is often used. The result of this is a large number of potentially noisy patterns which are deemed strong or interesting according to the data mining process. A personalised and interactive process is essential to support users in identifying the so-called nuggets of knowledge that are embedded in the discovered patterns.

Visualisation effectively communicates complex ideas and experimental results across disciplines. A significant number of general purpose data visualisation systems have been proposed [12, 34, 81]. These are generic approaches that enable users to load data, cluster it and visualise it in low-dimensional projections. These are limited by their generality, the need to understand algorithms and computer programming and are not designed for searching, exploring and visualising the geographic distribution of mined patterns. Pattern mining researchers have developed some visualisation tools, however, few of these have been proposed for co-location patterns [21], and there is no work in the literature on visualising spatial contrast nor common sets discovered in data mining.

In general, the AI research into pattern visualisation only offers a static perspective on the discoveries; specifically, the user does not have the opportunity to interactively produce the visualisations that are relevant to them in various levels of abstractions. Recent application of data science has noted the importance of interactive and exploratory tools for knowledge discovery and decision support in genetic data and temporal medical data [55, 72].

3.2 Methodology

In this section, we present the datasets, software and evaluation process used in our work.

3.2.1 Data and Preprocessing

The adverse birth outcomes datasets used in this research were acquired from the Alberta Perinatal Health Program (APHP)[3] and the Canadian Neonatal Network (CNN).[4] In each dataset, there are three main adverse birth outcomes: (1) Preterm birth (PTB)—a birth that takes place more than three weeks before the baby is due; (2) Low birth weight at term (LBW)—a birth in which the weight of the baby is less than 2500 g and the gestational age is on or above 37 weeks and (3) Small for gestational age (SGA)—birth in which the baby's weight is in the lower 10th percentile for the gestational age according to Kramer's Canada-wide statistics [42].

The APHP database is a rich dataset including mother's geolocated reported residence by postal code for all live births during the period of 2006–2012 for the province of Alberta, Canada. Specifically, the dataset contains the birth outcome (non-ABO, PTB, SGA, LBW) mother's residence location of 333,247 births. The distribution of the adverse birth outcomes in this dataset is as follows: (1) PTB 22,733 cases; (2) LBW 5,485 cases and (3) SGA 29,679 cases. The CNN data is collected from 19 Census Metropolitan Areas (CMAs) in all provinces across Canada through Neonatal Intensive Care Units (NICUs). This contains mothers admitted to NICUs during the time period of 2006–2010. In particular, the CNN dataset has the geolocated reported residences of 32,836 mothers along with their birth outcomes. The distribution of the adverse birth outcomes in this dataset is as follows: (1) PTB 17261 cases; (2) LBW 1476 cases and (3) SGA 5465 cases.

The industrial air pollutant emissions data were accessed via the National Pollutant Release Inventory (NPRI) of Canada for the time period of 2006–2012. The emissions dataset includes estimates of yearly releases of 136 industrial chemicals.

The NPRI map in Fig. 2 shows the distribution of the 6279 industrial facilities for the province of Alberta. The subsequent maps help demonstrate the distribution

[3] Alberta Perinatal Health Program. http://aphp.dapasoft.com/Lists/HTMLPages/index.aspx.

[4] http://www.canadianneonatalnetwork.org/portal/.

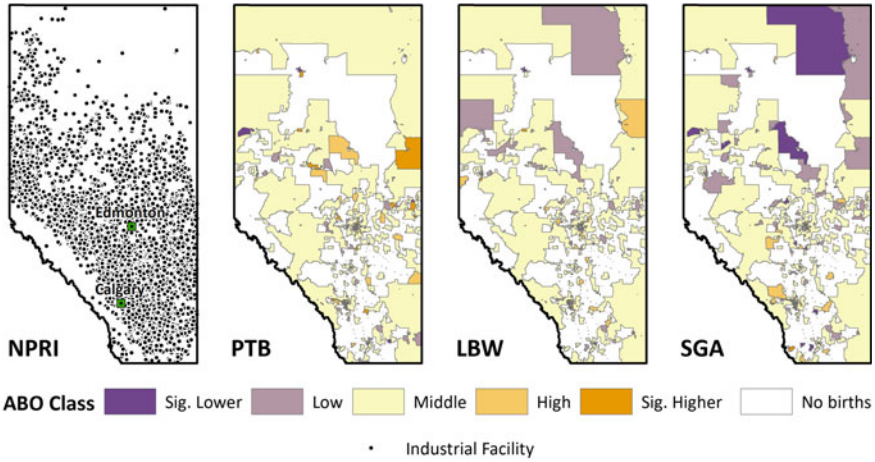


Fig. 2 From left to right, this figure shows the distribution of NPRI sites and the rates of PTB, LBW and SGA for in Alberta during the study period

of births by ABO. To protect individual privacy, the actual locations of residences cannot be shown; therefore, the categories are based on smoothed Bayesian rates that indicate areas of relatively lower (purple) and higher (orange) than the average (yellow) provincial rates of PTB, LBW and SGA. The maps were made available in the Web Mercator projection for knowledge users to access in the visualisation tool.

3.2.2 Transactionization

The above adverse birth outcomes and chemical emission datasets were integrated and tabulated via the transactionization process [39, 50]. To determine the overlapping regions of chemicals and births, we utilised historic weather data from Environment Canada and the Alberta Agriculture weather stations to simulate the atmospheric transportation of airborne chemicals from their point sources. We generated the dispersion region of an air pollutant from an emission point (facility) as a circular buffer where the centre was the emission point, and the radius was defined based on the amount of chemical released. To better reflect the dispersion area, we transform the circular region into an elliptical buffer region based on the average wind speed and direction. The lengths of the major axis 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 [$r = \ln(\text{amounts})$]; ν was the wind speed, and γ was the stretching coefficient (=0.3). Detailed information about this process has been published by [39].

As a surrogate of the maternal mobility range during pregnancy, a 5 km radius circles centred on the postal code location of the maternal residence are defined.

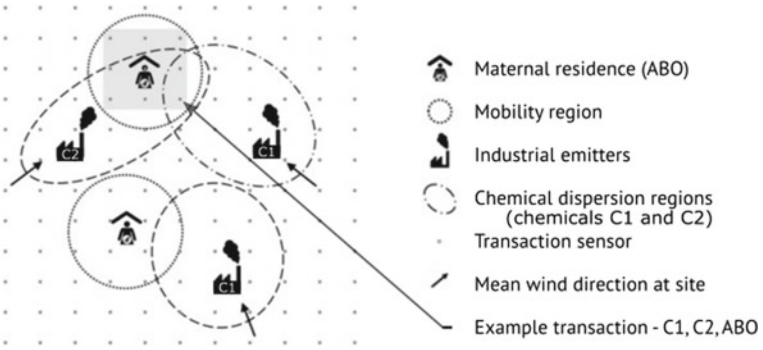

C2
C1
C1
Maternal residence (ABO)
Mobility region
Industrial emitters
Chemical dispersion regions (chemicals C1 and C2)
Transaction sensor
Mean wind direction at site
Example transaction - C1, C2, ABO

Fig. 3 This figure is recreated from [70]. It presents the process of transactionization. The geospatial region includes the maternal residences and mobility buffers, along with the chemical emission sources and downwind dispersion areas. Transactions record the birth outcomes and chemical occurrences for each grid point on the map according to the overlapping mobility and dispersion regions

This is overlaid with the region of interest (map) with a set of uniformly distributed grid points (1-km grid). This is illustrated in Fig. 3. Each grid point recorded the occurrence or absence of each event (ABO or non-ABO) and each industrial chemical at its location. Thus, an example transaction for a grid point is {SGA = True, LBW = False, ... , benzene = True, chlorine = False, PM = True, ...}. Each grid point is added to the transaction database for co-location pattern mining. As an example, the grid point highlighted in the figure records the co-occurrence of chemical C1, C2 and the ABO. Furthermore, through a transaction aggregation process, this algorithm also captures more complex scenarios where the mother was exposed to multiple chemicals, each with non-overlapping buffer regions.

3.2.3 Data Mining with AGT-Fisher

After the transaction dataset of birth outcomes is created, pattern mining with the Kingfisher dependency rule search technique is applied. Our previous work [39] demonstrated that the Kingfisher algorithm [33] finds non-redundant statistically significant co-location patterns between chemical mixtures and ABO. Kingfisher judges the statistical significance of the association between chemical mixtures and ABO using Fisher's exact test.

The Kingfisher algorithm enumerates 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 set → ABO or chemical set → non-ABO, where the pattern satisfied a *p*-value threshold. A *p*-value cut-off of 0.05 is used in this work. As previously stated, a common challenge in data mining is that the list of discovered associations remains large (i.e., hundreds). Moreover, it is highly likely that only a small subset of these patterns is of interest to knowledge users. As a result, a subsequent interactive and exploratory processes is needed to enable end-users to understand and isolate the most valuable knowledge in the discovered associations.

3.2.4 VizAR

VizAR is a formalisation for personalised rule identification that enables users to interact with, explore and visualise the discovered co-location patterns at three levels of abstraction:

1. Overview level
2. Pattern level
3. Instance level.

System Architecture: The architecture of VizAR is presented in Fig. 4. VizAR communicates with a central database that stores the previously mined patterns. By mining and storing the patterns in advance, we achieve three desirable outcomes: (a) the technical complexity of data mining is removed from the end-user, (b) the user experience is separated from the time complexity of the data mining, and (c) patient data is securely kept offline. In addition to the patterns, the pattern database includes the anonymised transactions on which the patterns were mined, the corresponding measures of interestingness and meta-data that enables geo-spatial visualisation and exploration of the discovered patterns. VizAR interacts with cloud services to access various kinds of resources such as maps and customised context on adverse birth outcome rates and socio-economic status.

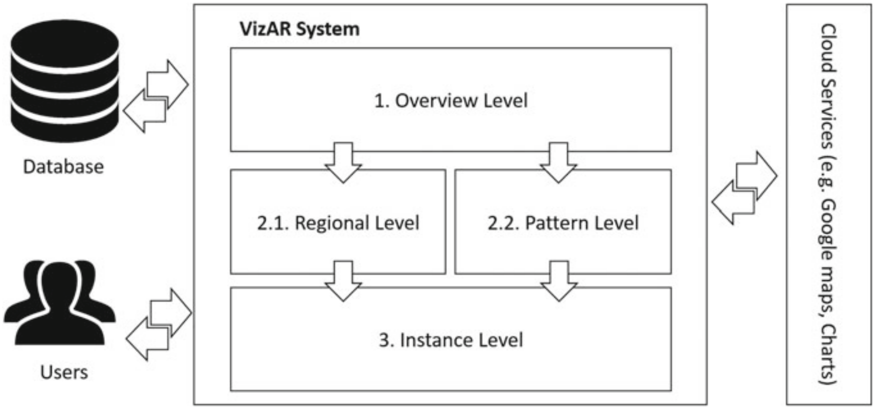


Fig. 4 System architecture of the VizAR framework

VizAR Data Views: The overview level has two components. Users can visualise the geo-spatial distribution of all of the co-location patterns and/or search, sort and filter for a subset of co-location patterns.

The geo-spatial distribution is depicted in the form of an interactive bubble chart in Fig. 5. It enables policymakers and other users to identify common or contrasting trends in sub-regions (CMAs). The CMAs are listed on the y-axis with spacing approximately scaled to the distance between their physical locations. The x-axis specifies the unique identifier of each discovered co-location pattern.

The occurrence of a pattern at a CMA is represented by a circle. In cases where an individual pattern (rule x) is discovered at multiple CMAs, circles are drawn at the intersection of the rule ID on the x-axis and the CMA on the y-axis. For each CMA, the size of the circle indicates the *support* in the dataset for that rule at that CMA. The colour indicates the statistical significance of the pattern at the CMA in terms of Fisher's exact test (log(p_F)) [39].

EX1 in Fig. 5 illustrates an example of contrasting regions (cities in this case) that can be discovered with this view. In this example, Toronto and Moncton are identified as contrasting regions because Toronto has significantly more co-location patterns associated with it than Moncton does. In a similar manner, users can easily identify regions that have association rules in common, such as Toronto and Montreal. We refer to these as geo-spatially common regions, which occur when regions have similar sets of rules. Policymakers can, for example, use this view to identify CMAs with similar issues and initiate working groups to develop focused research on specific chemicals and mixtures in order to support future development of solutions.

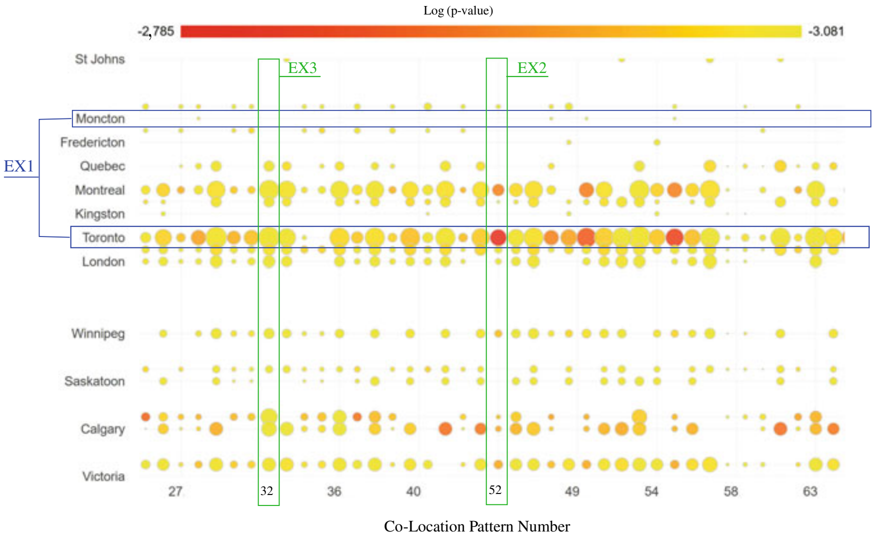
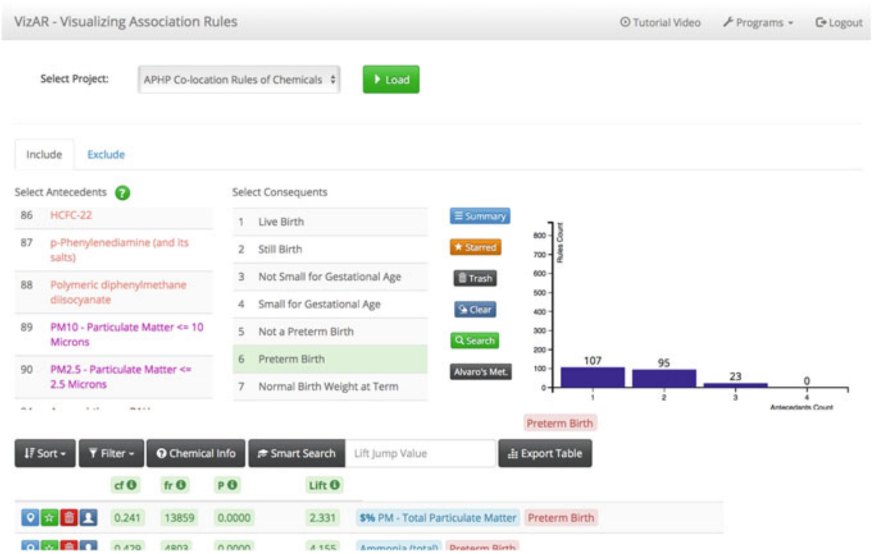


Fig. 5 Regional level visualisation: Bubble sizes represent the support of a rule in a particular spatial region, and the color code represents the log(p_F) range



Select Antecedents

86	HCFC-22
87	p-Phenylenediamine (and its salts)
88	Polymeric diphenylmethane diisocyanate
89	PM10 - Particulate Matter <= 10 Microns
90	PM2.5 - Particulate Matter <= 2.5 Microns

Select Consequents

1	Live Birth
2	Still Birth
3	Not Small for Gestational Age
4	Small for Gestational Age
5	Not a Preterm Birth
6	Preterm Birth
7	Normal Birth Weight at Term

cf	fr	P	Lift		
0.241	13859	0.0000	2.331	$% PM - Total Particulate Matter	Preterm Birth
0.429	4803	0.0000	4.155	Ammonia (total)	Preterm Birth

Fig. 6 Search, sort and filter view. This enables the user to find patterns involving chemicals and birth outcomes related to their research

EX2 and EX3 in Fig. 5 demonstrate that users can discover common or contrasting patterns. For instance, EX3 depicts a pattern which is uniformly statistically significant in multiple sub-regions (i.e. common set), whereas EX2 has divergent degrees of significance in different CMAs. Specifically, pattern 52 (i.e. EX2) has strong support and significance in Toronto and weak support elsewhere; thus, it is a contrasting pattern. This can lead policymakers to address the question, why is it prominent in Toronto and nowhere else? Alternatively, pattern 32 is a geo-spatially common pattern with significance and support similar across many CMAs.

The tabular frame in the overview level enables users to find and analyse the occurrence of patterns involving subsets of chemicals and/or adverse birth outcomes. It is shown in Fig. 6. Users can use this feature to reduce the scope of the bubble chart prior to analysis or drill down into the distribution of a specific pattern. In addition, summary statistics describing the number of patterns meeting a search requirement are produced. This includes the bar chart showing the number of patterns of each size that were found. Here, the pattern size refers to the number of chemicals involved. This is depicted on the right-hand side of the view.

Pattern Level Visualisation: Users can drill down to the pattern level view, depicted in Fig. 7, by selecting a pattern of interest at the overview level. This view presents a map of the entire geo-spatial region of interest annotated with the existence of the selected pattern. This gives a perspective on distribution of the pattern of interest across the CMAs in Canada. Once again, the occurrence of the pattern is depicted as a circle, where the support and significance are represented by size and colour. The example in this figure presents another way of identifying geo-spatially common and

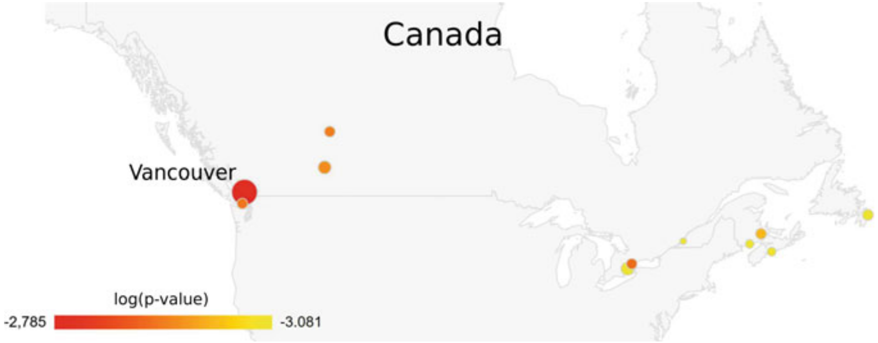


Fig. 7 Visualising a pattern's prevalence/significance in different CMA regions. Size of the bubbles represents the support for the pattern in a particular region, and colour code represents the log(p_F) (i.e. log of the Fisher's p-value) range

contrasting pattern. In this case, the pattern has much greater support and significance in Vancouver than the other CMAs (i.e. a contrasting pattern/set).

In addition to the map-based analysis, strategies for pattern level analysis based on wind rose plots are provided. A wind rose is typically used to visualise the relative frequency of wind speed at a specific location. It is used here to emphasise the disagreements in the support or significance of a particular pattern across the CMAs.

A spatial-temporal perspective on the patterns is also possible and very useful. This can be achieved using the wind rose plot. Figure 8, for example, demonstrates the visualisation of a pattern across spatial regions in different months. With this visualisation, users can, for instance, discover temporal changes and population shifts leading to a change in the distribution of the pattern.

Instance Level Visualisation: The lowest level of abstraction is formulated as the instance level view. It focuses on exploring a specific pattern in a specific CMA. This view presents the individual occurrences of the pattern on an interactive map. Figure 9 depicts the distribution of the occurrence of the pattern ($PM_{2.5}$, Methyl ethyl keyton, Xylene) $\rightarrow$ PTB in the Edmonton, Canada CMA. Users can zoom in and out on the map in order to gain perspectives on the distribution of the pattern down to the neighbourhood level. In addition, the view allows users to overlay other pertinent information, such as location of the emitting facilities, the interpolated dispersion regions of the chemicals, along birth outcome rates and socio-economic information broken down by denomination area. These help users to better understand the population under study.

4 Results

In order to demonstrate the efficacy of our framework, we present a summary of patterns identified through VizAR by our user-base. In addition, we describe two

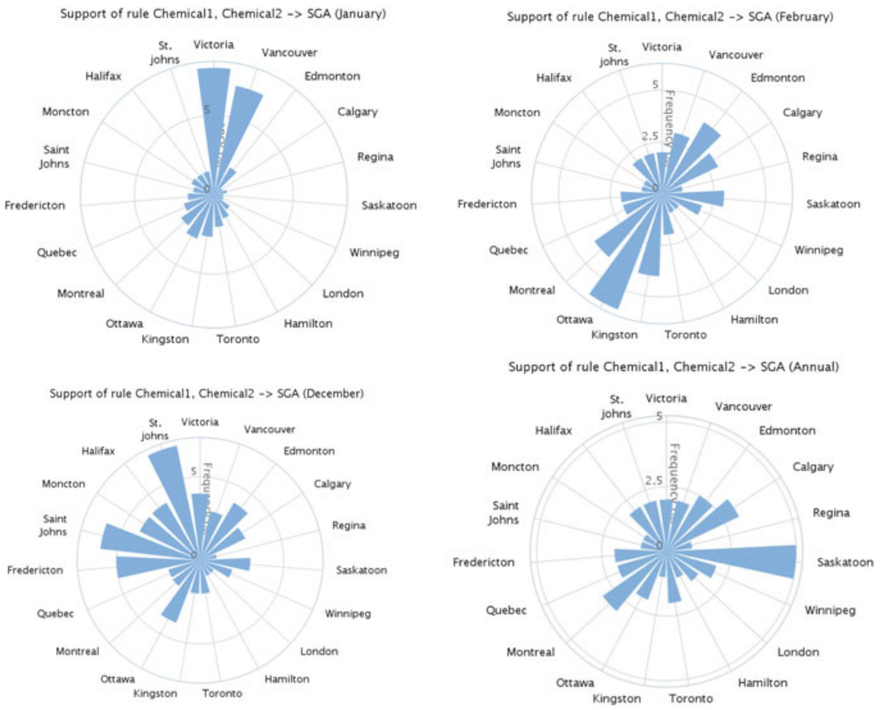
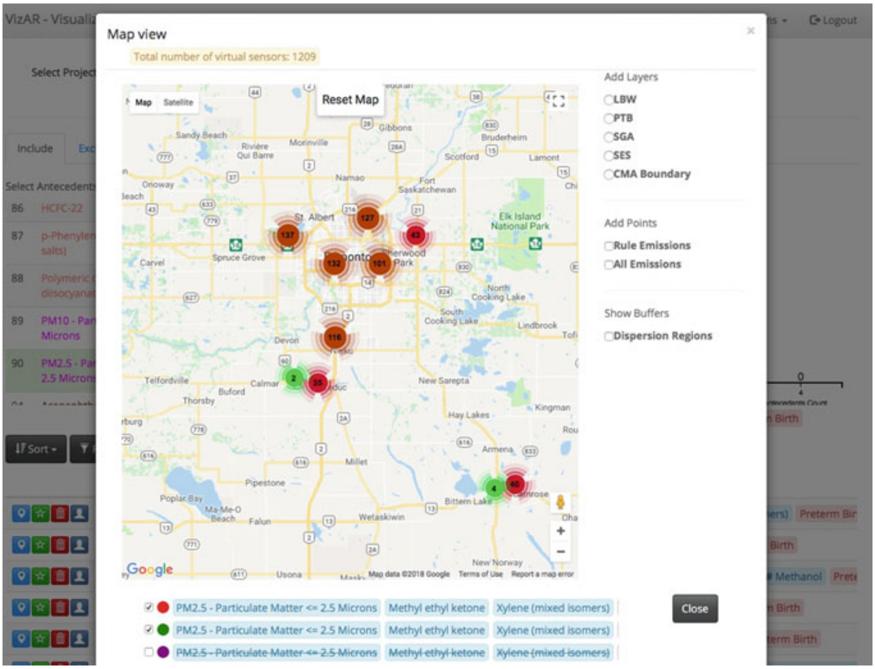


Fig. 8 Demonstration of spatial-temporal analysis by geography and birth month with wind rose plots. This shows a mock-up of the support distribution of contrast sets for January (top left), February (top right), December (bottom left) and annual average (bottom right)

exploration strategies employed by our users and conduct a thorough epidemiological assessment of one of the identified patterns. This is done by calculating the odds ratios of exposure and the outcome [70].

4.1 Identified Patterns of Interest

The user-base includes researchers in environmental health, epidemiology, neonatology, paediatrics and public health. Users were trained to use VizAR and given an opportunity to use it to explore the co-location patterns discovered by our AGT-Fisher algorithm on the datasets. A summary of the chemical mixtures of interest identified by the users via VizAR is provided in Table 1.



Map view

Total number of virtual sensors: 1209

Add Layers

○LBW
○PTB
○SGA
○SES
○CMA Boundary

Add Points

□Rule Emissions
□All Emissions

Show Buffers

□Dispersion Regions

☑ ● PM2.5 - Particulate Matter <= 2.5 Microns | Methyl ethyl ketone | Xylene (mixed isomers)
☑ ● PM2.5 - Particulate Matter <= 2.5 Microns | Methyl ethyl ketone | Xylene (mixed isomers)
□ ● ~~PM2.5 - Particulate Matter <= 2.5 Microns~~ | ~~Methyl ethyl ketone~~ | ~~Xylene (mixed isomers)~~

Fig. 9 Instances level visualising of the co-location pattern ($PM_{2.5}$, Methyl ethyl keytone, Xylene) → PTB in the Edmonton CMA. Green colour bubbles represent the places where only the antecedents exist (i.e. air pollutants), whereas the red bubbles represent the places where both the ABO and air pollutants coexist

Table 1 This table presents some chemical mixtures discovered to be associated ABOs via VizAR

Cmemicals	Outcome
Lead + Toluene	ABO
Lead + Xylene	ABO
Lead + Nitrogen dioxide + Particulate matter	ABO
Mercury + Phenanthrene	PTB
Metals + Polycyclic aromatic hydrocarbon	ABO
Toluene + Xylene + Methanol + Carbon monoxide	ABO
Ethylbenzene + Methyl isobutyl ketone	ABO
$PM_{2.5}$ + Methyl ethyl ketone + Xylene	PTB

The mixtures are either associated with general adverse birth outcomes (ABO), or a specific outcome, such as PTB

4.2 Discovery Techniques

Use-case 1 (geo-spatial exploration): This use-case is focused on identifying patterns of the form (chemical mixture) → ABO that have high lift values at multiple CMAs (geo-spatial common patterns). The discovery process analyses the bubble chart to find the rule with greatest significance for each CMA. The rules for each CMA are tabulated, and the most frequently occurring pattern is identified as a significant common pattern. In this case, it identifies that the pattern Lead → SGA to is the most significant common pattern of size one. It has the highest lift in 13 out of 19 CMAs. This indicates that the association between lead and SGA should be a significant question of interest in the majority of cities in our study area.

Use-case 2 (tabular search, sort and filter): The objective of this use-case is to efficiently find subsets of airborne chemical mixtures for which the exposed group has a significantly greater risk of having an adverse birth outcome than not having it. This requires searching and sorting to produce two ordered set P and N. P is a set of patterns $X \rightarrow A = a$ that is sorted according to lift, where the birth outcome is always $A = a$, and X is a set of chemicals. Alternatively, N is a sorted set of negative patterns $Y \rightarrow A = \overline{a}$. A score of the exposure risk is calculated from these sets using the lift ratio:

$$LR(X, A = a) = \text{lift}(X, A = a)/\text{lift}(X, A = \overline{a}) \tag{1}$$

The lift ratio utilises the intuition that all of the mothers in the CMA are exposed to the chemicals X. Thus, the larger the lift ratio, the more significant the association between the exposure and the ABO. The lift ratio is calculated for each pattern in P_i, $\{P_i : X_i \rightarrow A = a\} \in P$ that has a corresponding pattern N_j, $\{N_j : Y_j \rightarrow A = \overline{a}\} \in N$, such that the chemical mixtures are equivalent, $X_i = Y_j$.

Five patterns were discovered using this method, and the pattern ($PM_{2.5}$, Methyl ethyl ketone, Xylene) → PTB was selected to evaluate using odds ratio, which is a standard metric for risk assessment in epidemiology. The odds ratio is defined as the ratio between odds of adverse birth outcome among exposed versus unexposed groups. Thus, an odds ratio greater than 1 indicates a positive relationship between the exposure and the adverse birth outcomes. This pattern has an odds ratio of 1.14, which means that the exposed group is at greater risk than the unexposed group.

Figure 10 gives a relative perspective on the significance of this pattern. It shows the odds ratios, with 95% confidence intervals[5] for smoking and PTB, low socio-economic status (SES) and PTB, our discovered chemical mixture and PTB, along with the combination of all three (smoking, low SES status, rule 1) and PTB. The odds ratios for smoking and SES were calculated using maternal data form the APHP and Census data. In addition to showing that this chemical mixture poses a similar risk as other known factors, it demonstrates that the combination of the chemicals, smoking

[5] Adjusting for maternal confounders including smoking, substance use, past-preterm, mothers' age, socio-economic status, etc.

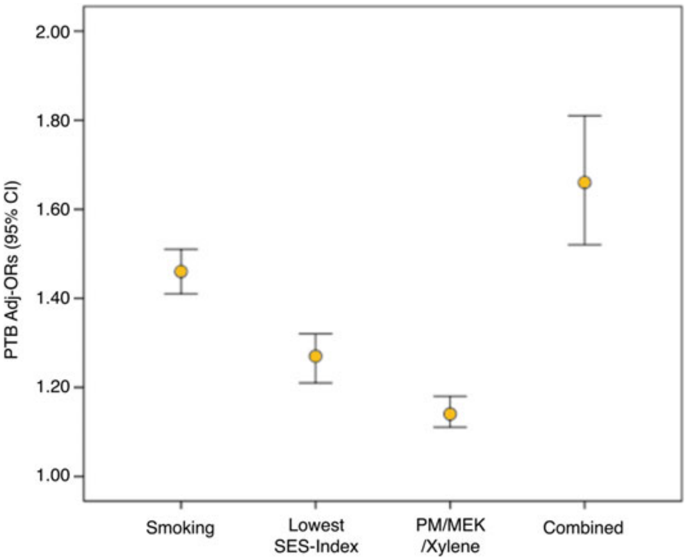


Fig. 10 Comparison of odds ratio (adjusted for relevant maternal confounders and socio-economic status) for an association discovered for preterm birth with other risk factors

and low SES status poses an even greater risk than the individual components. This finding has, in fact, inspired future work and a grant proposal.

5 Discussion

There is a growing body of literature and practical examples that demonstrate the great potential for AI to support the advancement of environmental health. Nonetheless, numerous challenges exist, such as access to a sufficient amount of high-quality data, how optimally pair AI with existing methods in environmental health, appropriate AI algorithm evaluation and parameter tuning methods and techniques to report results in manner that is understandable and reproducible by an interdisciplinary audience. These and related topics are discussed in the subsections below.

5.1 Pattern Filtering and Hypothesis Generation

As demonstrated by the DoMiNO case study, data mining is particularly powerful in contexts involving mixtures of airborne chemicals. The number of patterns discovered by data mining methods, however, can be large and intractable for human analysis. As a result, pattern filtering and visualisations approaches are needed to reduce the volume of discovered patterns. In the DoMiNO project, the lift ratio is utilised to filter the output of the AGT-Fisher algorithm, and hypothesis generation is performed via the interactive visualisation provide by VizAR. The authors in [77] developed two post-pruning criteria to filter the output of the basic Apriori algorithm.

The exploratory nature of data science implies that users are often looking for new insights without a-priori knowledge of the form that the patterns might take. When filtering is applied, it is important to recognise that it risks removing associations between rare, but critical, mixtures and outcomes [62]. Researchers must be careful to achieve the satisfactory balance between reducing the number of patterns and maintaining good sensitivity. The combination of filtering and interactive visualisation can facilitate a better in this respect. Viewing data mining results with GIS tools has also been demonstrated to be a effective way to discover meaningful patterns [63]. Nonetheless, additional research on best practices for pattern filtering and hypothesis generation in the context of environment health is needed.

Because the AI algorithms generally find associations rather than causation, they are better suited to serve as the first step in the hypothesis generation process. The authors in [75] demonstrate AI coupled with traditional methods to narrow the search space. The benefit of such a combined system is that the AI can be applied to high-dimensional, continuous exposure variables, and traditional epidemiological methods control for confounding, assess effect size, investigate various contrasting exposures and identify chemical mixtures of interest.

5.2 Data

Exposome [83] and other ambitious projects are expanding the size and scope of what is traditionally studied in environmental health. To support the characterisation of the breadth of exposures that humans encounter from birth to death necessitates the design and evaluation of novel AI methods for exceptionally high-dimensional spatial-temporal datasets. Advancements, such as those seen in natural language processing with LSTM and transformer networks, are needed in order to discover critical links between events with significant temporal separation.

Regardless of the above-mentioned efforts, the authors in [62] note that the publicly available data remains a significantly limited. Challenges with respect to data access include the cost and complexity of pollution monitoring and dispersion modelling, along with inconsistent collection and privacy concerns related to health records. Whilst the number of potential exposure combinations is immense, the pollution monitoring and health outcome data remain sparse. As a result, the authors claim that the current data may not allow for reproducible findings. New research focused on the application of AI to small, high-dimensional and sparse environmental health data is needed.

Moreover, the quality of the available data is an issue that requires attention. The accuracy of the available data can be compromised on many fronts. This includes due to human error and the accuracy of sensors or the dispersion models used. In addition, the imperfect output of AI algorithms is used as the input to subsequent AI models. This can add a degree of uncertainty to the data that most AI methods cannot account for [74]. AI algorithms that are robust to noise and provide well-calibrated confidence scores will be a great use.

5.3 Robustness and Validity

Unlike traditional statistics, the focus in much of AI is on designing and developing accurate predictive models and discovering frequent, but unknown, patterns. It is critical that collaborators in interdisciplinary application, such as environmental health, understand the implicit assumptions and objectives being optimised by the AI algorithms used (e.g. finding associations versus causal relationships). In many cases, terminology may be used or understood differently between fields. For collaborations to be successful, issues of this nature should be identified in advance and clarified in subsequent publications.

As discussed in [84], interdisciplinary collaborations between AI and environmental health researchers can serve as a gateway to new results and discoveries. These collaborations require that the participants commit time to relationship building, continuous learning and engagement in order to mitigate conflicts and misunderstandings. DoMiNO utilised an iterative process of learning and familiarisation to establish a common ground with regards to data mining methodologies and terminologies. This was found to increase the likelihood of success by providing collaborators from across disciplines with the skill set necessary to proactively participate in the design and undertaking of the data mining process.

In order to promote robust and appropriate use of AI in environmental health, it is advised that practitioners explicitly state the goal of the study in advance, explain why AI is needed and what the assumptions and risks are. Simulation studies and analyses of the AI on artificial datasets that replicate key properties of the target domain are an excellent means of building trust in and understanding of the proposed method. The authors in [44] used a simulated study to assess boosted regression trees' ability to detect relationships between chemical mixtures and metabolic syndrome. This serves to simplify the identification of the limitation of the method, evaluate its robustness to training sets size, noise and correlated exposures.

Results of the AI algorithms and the hypotheses generated from them ought to be considered in the context of the representativeness of the data used. Much like science and society in general, it has been shown that the results of AI algorithms suffer from bias [11, 15]. Recent work has also discussed racism in algorithms deployed in health care [60]. Whilst the representativeness of the data is a major point of consideration in environmental health, it is often overlooked in AI where the academic focus has typically been on theoretical considerations of algorithmic learning. It is only now becoming a critical point of consideration in academic and industrial AI [88]. In the context of AI applied to environmental health, spatial variability in exposure profiles, demographics and contextual characteristics of the subjects in the data must be considered.

5.4 Transparency and Trust

Transparency and trust play an important role when it comes to health and medical applications. The most powerful AI algorithms tend to be complex and are less transparent. This is particularly the case for modern end-to-end deep learning system. Hence, achieving transparency and maintaining trust whilst building a successful AI system is a challenging task, especially in an interdisciplinary setting. In part, the DoMiNO project accomplished this through dialog and mutual learning, but also by facilitating a human-machine interactive process where end-users actively become part of the knowledge discovery process with VizAR. Rather than passively consuming patterns/knowledge provided by the algorithm, the users interactively explored them to understand their foundation and meaningfulness.

5.5 Deep Learning

Artificial neural networks date back to the 1960s. As a result of significant improvements in computing power and dataset size, along with refinements in the learning algorithms, the modern incarnation of artificial neural networks (deep learning) can achieve human-level performance in a wide variety of applications including health [30, 36].

In environmental health, deep learning algorithms designed for object recognition tasks, such as convolutional neural networks (CNNs), have a great potential [43]. Supported by the growing availability of ground- and satellite-based imagery, CNNs provide the potential to simplify and improve large-scale pollution modelling and air quality prediction [82]. A large portion of environmental health data, including that from air pollution senors and medical records, is sequential. Like image recognition, deep learning has made significant breakthroughs in modelling and predicting sequential data, such as natural language [32, 78]. With the growing availability of sequential environmental health data, deep learning architectures, such as LSTMs and transformers, have a great potential to improve the predictive performance beyond the current standard.

Missing data is a common problem in both statistics and AI. In general, it may be handled by removing records with missing values or filling the missing values with estimates and data imputation [5]. However, domain-specific approaches may be devised that produce better results. Missing values, for example, occur in AOD data due to cloud cover and other atmospheric conditions. In [14], the authors addressed this by training two deep learning models, one with and one without AOD data. In other settings, however, training two models may not provide satisfactory performance in all conditions. Deep generative networks, such as generative adversarial networks (GAN) [31], can serve as more powerful data imputation and augmentation methods [37, 47].

Other important challenges in environmental health relate to limited, sparse and class imbalanced data. This includes the lack of pollution data from rural areas and in marginalised and low-income communities. As a result, there is a dearth of knowledge about health issues that are specific to these communities. It is critical that the growing potential of AI in environmental health is utilised to benefit these communities that have traditional been under-served. In addition to new algorithms and data sources, this will require working with communities to better understand their environmental health wants and needs.

Learning from limited data is a challenge that transcends many deep learning applications. It is a quickly developing field of study that has generated a great amount of interest [1–3]. Some exemplary methods with potential in environmental health include data augmentation, transfer learning, domain adaption, few-shot learning and meta-learning. Data augmentation methods serve to correct for class imbalance and artificially inflate the number of samples from underrepresented populations [8, 9, 59]. Few-shot learning and meta-learning aim to utilise knowledge from earlier phases of training to quickly learning new predictive capabilities [27, 28]. In the context of environmental health, this offers the potential for the model to quickly adapt to new health outcomes and new prediction settings. Transfer learning and domain adaption, on the other hand, are techniques that enable models pre-trained one dataset to be quickly refit to new, but typically related, dataset. This can enable better generalisation in the transferred domain and faster learning [30]. A possible application is to develop models for cities with limited data by pre-training on data from cities with a large, representative network of air quality sensors.

6 Summary

Exposure to pollution in the environment is a major contributor to disease globally. There remains, however, a dearth of knowledge about the levels, distribution and types of airborne pollutants in the environment, along with how exposure to complex mixtures of airborne chemicals impacts health outcomes. Research in environmental health aims to monitor and understand factors in the environment that affect human health and disease. Recent collaborations between AI researchers and environmental health have demonstrated a great potential to help advance the science of air pollution epidemiology, urban planning and public policy.

In this chapter, we discussed AI in the context of environmental health related to air pollution. We outlined the importance of the field of study, the challenges that it currently faces and the opportunity for AI to contribute to the advancement of the field. In addition, we presented a case study on the DoMiNO project, which utilised

AI algorithms in combination with pattern visualisation via VizAR and traditional epidemiological analysis to generate hypotheses about which mixtures of airborne chemicals have the greatest impact on birth outcomes. Our results highlight both the great potential for AI in this field along with some interesting challenges for AI researchers to address in future work with environmental health researchers.

References

1. The 2nd learning from limited labeled data (lld) workshop. https://lld-workshop.github.io/. Accessed: 2021-03-29
2. From shallow to deep: Overcoming limited and adverse data. https://s2d-olad.github.io. Accessed: 2021-03-29
3. Workshop on meta-learning (metalearn 2020). https://meta-learn.github.io/2020/. Accessed: 2021-03-29
4. Agarwal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: Proc. of the 20th VLDB Conference, pp. 487–499 (1994)
5. Aggarwal, C.C.: Data mining: the textbook. Springer (2015)
6. Austin, E., Coull, B., Thomas, D., Koutrakis, P.: A framework for identifying distinct multipollutant profiles in air pollution data. Environment international **45**, 112–121 (2012)
7. Bell, S.M., Edwards, S.W.: Identification and prioritization of relationships between environmental stressors and adverse human health impacts. Environmental health perspectives **123**(11), 1193–1199 (2015)
8. Bellinger, C., Corizzo, R., Japkowicz, N.: Remix: Calibrated resampling for class imbalance in deep learning. arXiv preprint arXiv:2012.02312 (2020)
9. Bellinger, C., Drummond, C., Japkowicz, N.: Manifold-based synthetic oversampling with manifold conformance estimation. Machine Learning **107**(3), 605–637 (2018)
10. Bellinger, C., Jabbar, M.S.M., Zaïane, O., Osornio-Vargas, A.: A systematic review of data mining and machine learning for air pollution epidemiology. BMC public health **17**(1), 1–19 (2017)
11. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 4356–4364 (2016)
12. Bostock, M., Heer, J.: Protovis: A graphical toolkit for visualization. IEEE transactions on visualization and computer graphics **15**(6) (2009)
13. Braun, J.M., Gennings, C., Hauser, R., Webster, T.F.: What can epidemiological studies tell us about the impact of chemical mixtures on human health? Environmental health perspectives **124**(1), A6–A9 (2016)
14. Brokamp, C., Jandarov, R., Hossain, M., Ryan, P.: Predicting daily urban fine particulate matter concentrations using a random forest model. Environmental science & technology **52**(7), 4173–4179 (2018)
15. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, pp. 77–91. PMLR (2018)
16. Chakma, A., Vizena, B., Cao, T., Lin, J., Zhang, J.: Image-based air quality analysis using deep convolutional neural network. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 3949–3952. IEEE (2017)
17. Challoner, A., Pilla, F., Gill, L.: Prediction of indoor air exposure from outdoor air quality using an artificial neural network model for inner city commercial buildings. International journal of environmental research and public health **12**(12), 15,233–15,253 (2015)

18. Chen, H.W., Tsai, C.T., She, C.W., Lin, Y.C., Chiang, C.F.: Exploring the background features of acidic and basic air pollutants around an industrial complex using data mining approach. Chemosphere **81**(10), 1358–1367 (2010)
19. Chen, M., Wang, P., Chen, Q., Wu, J., Chen, X.: A clustering algorithm for sample data based on environmental pollution characteristics. Atmospheric Environment **107**, 194–203 (2015)
20. Coker, E., Liverani, S., Ghosh, J.K., Jerrett, M., Beckerman, B., Li, A., Ritz, B., Molitor, J.: Multi-pollutant exposure profiles associated with term low birth weight in los angeles county. Environment international **91**, 1–13 (2016)
21. Desmier, E., Flouvat, F., Gay, D., Selmaoui-Folcher, N.: A clustering-based visualization of colocation patterns. In: Proceedings of the 15th Symposium on international database engineering & applications, pp. 70–78. ACM (2011)
22. Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M.B., Choirat, C., Koutrakis, P., Lyapustin, A., et al.: An ensemble-based model of pm2. 5 concentration across the contiguous united states with high spatiotemporal resolution. Environment international **130**, 104,909 (2019)
23. Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., Schwartz, J.: Assessing pm2. 5 exposures with high spatiotemporal resolution across the continental united states. Environmental science & technology **50**(9), 4712–4721 (2016)
24. Dominici, F., Peng, R.D., Barr, C.D., Bell, M.L.: Protecting human health from air pollution: shifting from a single-pollutant to a multi-pollutant approach. Epidemiology (Cambridge, Mass.) **21**(2), 187 (2010)
25. Dominici, F., Peng, R.D., Bell, M.L., Pham, L., McDermott, A., Zeger, S.L., Samet, J.M.: Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. Jama **295**(10), 1127–1134 (2006)
26. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in knowledge discovery and data mining, vol. 21. AAAI press Menlo Park (1996)
27. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE transactions on pattern analysis and machine intelligence **28**(4), 594–611 (2006)
28. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135. PMLR (2017)
29. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR) **38**(3), 9 (2006)
30. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT press Cambridge (2016)
31. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014), pp. 2672–2680 (2014)
32. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing, pp. 6645–6649. Ieee (2013)
33. Hämäläinen, W.: Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures. Knowledge and information systems **32**(2), 383–414 (2012)
34. Heer, J., Card, S.K., Landay, J.A.: Prefuse: a toolkit for interactive information visualization. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 421–430. ACM (2005)
35. Hermann, J., Schätzle, Z., Noé, F.: Deep-neural-network solution of the electronic schrödinger equation. Nature Chemistry **12**(10), 891–897 (2020)
36. Hinton, G.: Deep learning—a technology with the potential to transform health care. Jama **320**(11), 1101–1102 (2018)
37. Hu, X., Belle, J.H., Meng, X., Wildani, A., Waller, L.A., Strickland, M.J., Liu, Y.: Estimating pm2. 5 concentrations in the conterminous united states using the random forest approach. Environmental science & technology **51**(12), 6936–6944 (2017)

38. Jabbar, M., Zaïane, O., Osornio-Vargas, A.: Discovering spatial contrast and common sets with statistically significant co-location patterns. In: Proceedings of the Symposium on Applied Computing, pp. 796–803. ACM (2017)
39. Jabbar, M.S.M., Bellinger, C., Zaïane, O.R., Osornio-Vargas, A.: Discovering co-location patterns with aggregated spatial transactions and dependency rules. International Journal of Data Science and Analytics **5**(2), 137–154 (2018)
40. Jalali-Heravi, M., Zaïane, O.R.: A study on interestingness measures for associative classifiers. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 1039–1046. ACM (2010)
41. Johns, D.O., Stanek, L.W., Walker, K., Benromdhane, S., Hubbell, B., Ross, M., Devlin, R.B., Costa, D.L., Greenbaum, D.S.: Practical advancement of multipollutant scientific and risk assessment approaches for ambient air pollution. Environmental health perspectives **120**(9), 1238–1242 (2012)
42. Kramer, M.S., Platt, R.W., Wen, S.W., Joseph, K., Allen, A., Abrahamowicz, M., Blondel, B., Bréart, G., of the Canadian Perinatal Surveillance System, F.H.S.G., et al.: A new and improved population-based canadian reference for birth weight for gestational age. Pediatrics **108**(2), e35–e35 (2001)
43. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
44. Lampa, E., Lind, L., Lind, P.M., Bornefalk-Hermansson, A.: The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. Environmental Health **13**(1), 1–17 (2014)
45. Landrigan, P.J., Fuller, R., Acosta, N.J., Adeyi, O., Arnold, R., Baldé, A.B., Bertollini, R., Bose-O'Reilly, S., Boufford, J.I., Breysse, P.N., et al.: The lancet commission on pollution and health. The lancet **391**(10119), 462–512 (2018)
46. Lary, D.J., Faruque, F.S., Malakar, N., Moore, A., Roscoe, B., Adams, Z.L., Eggelston, Y.: Estimating the global abundance of ground level presence of particulate matter (pm2. 5). Geospatial health pp. S611–S630 (2014)
47. Lee, H., Liu, Y., Coull, B., Schwartz, J., Koutrakis, P.: A novel calibration approach of modis aod data to predict pm 2.5 concentrations. Atmospheric Chemistry and Physics **11**(15), 7991–8002 (2011)
48. Lee, J.G., Kang, M.: Geospatial big data: challenges and opportunities. Big Data Research **2**(2), 74–81 (2015)
49. Lewis, A.C., Lee, J.D., Edwards, P.M., Shaw, M.D., Evans, M.J., Moller, S.J., Smith, K.R., Buckley, J.W., Ellis, M., Gillot, S.R., et al.: Evaluating the performance of low cost chemical sensors for air pollution research. Faraday discussions **189**, 85–103 (2016)
50. Li, J., Zaïane, O.R., Osornio-Vargas, A.: Discovering statistically significant co-location rules in datasets with extended spatial objects. In: International Conference on Data Warehousing and Knowledge Discovery, pp. 124–135. Springer (2014)
51. Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T.: Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental pollution **231**, 997–1004 (2017)
52. Li, Y., Huang, J., Luo, J.: Using user generated online photos to estimate and monitor air pollution in major cities. In: Proceedings of the 7th International Conference on Internet Multimedia Computing and Service, pp. 1–5 (2015)
53. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
54. Liu, C., Tsow, F., Zou, Y., Tao, N.: Particle pollution estimation based on image analysis. PloS one **11**(2), e0145,955 (2016)
55. Ltifi, H., Ben Mohamed, E., ben Ayed, M.: Interactive visual knowledge discovery from data-based temporal decision support system. Information Visualization **15**(1), 31–50 (2016)
56. Manrai, A.K., Cui, Y., Bushel, P.R., Hall, M., Karakitsios, S., Mattingly, C.J., Ritchie, M., Schmitt, C., Sarigiannis, D.A., Thomas, D.C., et al.: Informatics and data analytics to support

exposome-based discovery for public health. Annual review of public health **38**, 279–294 (2017)

57. Mauderly, J.L., Burnett, R.T., Castillejos, M., Özkaynak, H., Samet, J.M., Stieb, D.M., Vedal, S., Wyzga, R.E.: Is the air pollution health research community prepared to support a multipollutant air quality management framework? Inhalation toxicology **22**(sup1), 1–19 (2010)
58. Molitor, J., Coker, E., Jerrett, M., Ritz, B., Li, A.: Part 3. modeling of multipollutant profiles and spatially varying health effects with applications to indicators of adverse birth outcomes. Research report (Health Effects Institute) (183 Pt 3), 3–47 (2016)
59. Mullick, S.S., Datta, S., Das, S.: Generative adversarial minority oversampling. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1695–1704 (2019)
60. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
61. Oskar, S., Stingone, J.A.: Machine learning within studies of early-life environmental exposures and child health: Review of the current literature and discussion of next steps. Current Environmental Health Reports pp. 1–15 (2020)
62. Patel, C.J., Manrai, A.K.: Development of exposome correlation globes to map out environment-wide associations. In: Pacific Symposium on Biocomputing Co-Chairs, pp. 231–242. World Scientific (2014)
63. Pearce, J.L., Waller, L.A., Sarnat, S.E., Chang, H.H., Klein, M., Mulholland, J.A., Tolbert, P.E.: Characterizing the spatial distribution of multiple pollutants and populations at risk in atlanta, georgia. Spatial and spatio-temporal epidemiology **18**, 13–23 (2016)
64. Rajapakse, N., Silva, E., Kortenkamp, A.: Combining xenoestrogens at levels below individual no-observed-effect concentrations dramatically enhances steroid hormone action. Environmental health perspectives **110**(9), 917–921 (2002)
65. Ram, S., Zhang, W., Williams, M., Pengetnze, Y.: Predicting asthma-related emergency department visits using big data. IEEE journal of biomedical and health informatics **19**(4), 1216–1223 (2015)
66. Reid, C.E., Jerrett, M., Petersen, M.L., Pfister, G.G., Morefield, P.E., Tager, I.B., Raffuse, S.M., Balmes, J.R.: Spatiotemporal prediction of fine particulate matter during the 2008 northern california wildfires using machine learning. Environmental science & technology **49**(6), 3887–3896 (2015)
67. Represa, N.S., Fernández-Sarría, A., Porta, A., Palomar-Vázquez, J.: Data mining paradigm in the study of air quality. Environmental Processes **7**(1), 1–21 (2020)
68. SCHER (Scientific Committee on Health and Environmental Risks): Toxicity and assessment of chemical mixtures (2012). https://doi.org/10.2772/21444. https://ec.europa.eu/health/sites/health/files/scientific_committees/environmental_risks/docs/scher_o_155.pdf
69. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A., et al.: Improved protein structure prediction using potentials from deep learning. Nature **577**(7792), 706–710 (2020)
70. Serrano-Lomelin, J., Nielsen, C.C., Jabbar, M.S.M., Wine, O., Bellinger, C., Villeneuve, P.J., Stieb, D., Aelicks, N., Aziz, K., Buka, I., et al.: Interdisciplinary-driven hypotheses on spatial associations of mixtures of industrial air pollutants with adverse birth outcomes. Environment international **131**, 104,972 (2019)
71. Silva, E., Rajapakse, N., Kortenkamp, A.: Something from "nothing"- eight weak estrogenic chemicals combined at concentrations below noecs produce significant mixture effects. Environmental science & technology **36**(8), 1751–1756 (2002)
72. Simon, S., Mittelstädt, S., Kwon, B.C., Stoffel, A., Landstorfer, R., Neuhaus, K., Mühlig, A., Scherer, S., Keim, D.A.: Visexpress: Visual exploration of differential gene expression data. Information Visualization **16**(1), 48–73 (2017)
73. Singh, K.P., Gupta, S., Rai, P.: Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment **80**, 426–437 (2013)
74. Stingone, J.A., Pandey, O.P., Claudio, L., Pandey, G.: Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among us children. Environmental Pollution **230**, 730–740 (2017)

75. Sun, Z., Tao, Y., Li, S., Ferguson, K.K., Meeker, J.D., Park, S.K., Batterman, S.A., Mukherjee, B.: Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. Environmental Health **12**(1), 1–19 (2013)
76. Thurston, G.D., Spengler, J.D.: A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan boston. Atmospheric Environment (1967) **19**(1), 9–25 (1985)
77. Toti, G., Vilalta, R., Lindner, P., Lefer, B., Macias, C., Price, D.: Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. Artificial intelligence in medicine **74**, 44–52 (2016)
78. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
79. VoPham, T., Hart, J.E., Laden, F., Chiang, Y.Y.: Emerging trends in geospatial artificial intelligence (geoai): potential applications for environmental epidemiology. Environmental Health **17**(1), 1–6 (2018)
80. Wang, H., Zhao, L.: A joint prevention and control mechanism for air pollution in the beijing-tianjin-hebei region in china based on long-term and massive data mining of pollutant concentration. Atmospheric Environment **174**, 25–42 (2018)
81. Ward, M.O.: Xmdvtool: Integrating multiple methods for visualizing multivariate data. In: Proceedings of the Conference on Visualization'94, pp. 326–333. IEEE Computer Society Press (1994)
82. Weichenthal, S., Hatzopoulou, M., Brauer, M.: A picture tells a thousand... exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. Environment international **122**, 3–10 (2019)
83. Wild, C.P.: The exposome: from concept to utility. International journal of epidemiology **41**(1), 24–32 (2012)
84. Wine, O., Zaiane, O.R., Osornio Vargas, A.R.: A collaborative research exploration of pollutant mixtures and adverse birth outcomes by using innovative spatial data mining methods: The domino project. Challenges **10**(1), 25 (2019)
85. Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., Yoc, J.S.: A framework for discovering co-location patterns in data sets with extended spatial objects. In: Proceedings of the 2004 SIAM International Conference on Data Mining, pp. 78–89. SIAM (2004)
86. Xu, Y., Yang, W., Wang, J.: Air quality early-warning system for cities in china. Atmospheric Environment **148**, 239–257 (2017)
87. Zhang, L., Yang, G., Li, X.: Mining sequential patterns of pm2. 5 pollution between 338 cities in china. Journal of environmental management **262**, 110,341 (2020)
88. Zou, J., Schiebinger, L.: Ai can be sexist and racist—it's time to make it fair (2018)