Heliyon 10 (2024) e28537

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

An inferential spatiotemporal approach for knowledge synthesis to identify trends in public health research

Nicholas Grokhowsky

North Carolina State University, Raleigh, NC, USA

ARTICLE INFO

Keywords:
Research bias
Geographical bias
Research trends
Decision making
Knowledge synthesis

ABSTRACT

Background: Decisions follow patterns that are introduced by human perception. Research and development (R&D) are influenced by these patterns. Furthermore, R&D publications can represent repetitive attempts to solve similar, or the same problems. Literature reviews serve as an important tool for identifying these trends, but they are time consuming. The time commitment of a literature review can be reduced by using a sample of research. This will allow an infinite population of research to be generalized. Additionally, spatiotemporal analysis is most appropriate for fields that follow time and geographic trends, such as public health. Also, using research locations to perform this analysis potentially captures the social return of R&D, as knowledge gained. As a result, an inferential spatiotemporal methodological framework is introduced to quickly identify research trends using public health research. This was applied to a childhood Pb exposure case study.

Methods: A body of more than 1000 childhood elevated blood lead (Pb) level (EBLL) research articles were used to extract publication years, research locations, and subtopics. These publications were grouped into research locations (i.e., U.S. states where research was conducted; not publication location) and averaged over years published (i.e., 29 years). Binary indicator variables were derived using the subtopics extracted and the periods identified in time trend analyses. Explanatory variables were used to conduct hypothesis testing. Significant variables were used to generalize the population of the annual average EBLL articles written per state.

Results: The range of the annual average of EBLL research articles by state was 0–1.7 articles, with a mean of 0.3 articles. Thirty-eight explanatory variables suggested a significant effect on research article production. These included temporal, sociodemographic, education, structure age, environmental, and economic variables. The strongest effect on research production for U.S. states came from the number of structures built before 1950. A predictive model was selected to generalize the population of articles using time-periods 1990-95, environmental subtopic, and structures built before 1950. The locations with the most research production for this topic were California and New York. The locations with the least research production for this topic were Alaska, Hawaii, Nevada, Wyoming, North Dakota, South Dakota, Mississippi, Delaware, and New Hampshire.

Conclusion: If the trend for R&D is to make fast decisions, more human bias will be introduced into the decision-making process. Analytical tools that enable researchers to identify trends and ask more questions about their field will mitigate these biases. This hypothesis testing and predictive modeling methodology provide researchers and other decision makers with analytical tools they can use to quickly identify research trends and narrow their field of research. Additionally, this

E-mail address: nickgrok@gmail.com.

https://doi.org/10.1016/j.heliyon.2024.e28537
Received 19 May 2023; Received in revised form 18 March 2024; Accepted 20 March 2024
Available online 26 March 2024

analysis potentially captures the *impact of discovered ideas*, as a social return spillover, for this topic.

1. Background

Decisions follow patterns caused by the amount of time available for the decision-making process and the magnitude of a decision's impact. Faster decisions and higher impact decisions have more biases than decisions that are made more slowly or are less impactful [1,2]. Many quick, high impact decisions are made in research and development (R&D), using large amounts of information (e.g., scientific journal articles). The same topics and subtopics are often researched repetitively, as confirmation of findings is a critical part of the scientific process. Additionally, the decisions to investigate these same research problems are influenced by previously published research. The previously published research typically reinforces knowledge and beliefs about the topic [3]. As a result, large quantities of research contain decision making trends. This concept provides a method for identifying the influences affecting researcher questions in specific fields of study.

Ecology provides one illustration of how research trends can be assessed using geographic distributions. For example, an ecologist studying breeding habits of the ruby throated hummingbird (i.e., *Archilochus colubris)* will likely focus on the ruby throated hummingbirds' breeding range, which is Central America to the eastern half of North America [4]. An unbiased research trend will have research distributed throughout the geographic breeding range with more emphasis on locations within the hummingbirds' most common breeding grounds. However, geographically biased research trends will have more of this research conducted in the hummingbirds' least common breeding grounds and minimal research elsewhere. This geographic bias can be caused by many types of bias, such as confirmation bias, anchoring bias, or a framing effect, to name a few. To emphasize this hypothetical example, a recent research paper analyzed the geographic distribution and species distribution of animal pollinator research publications [5]. The outcome was that more than 50% of the publications had research locations in five countries, 20% of the publications were specific to the bumble bee (i.e., *Bombus*), and 25% of the publications were specific to the honey bee (i.e., *Apis*). On a global scale, these statistics show a potential bias to study two of 200,000 animal pollinators [6] in five of 193 countries [7]. The researchers suggest the research is biased because of selection bias due to an easy-to-access data base on the animal pollinator topic [5]. These results highlight the importance for R&D decision makers to consider research trends before deciding on a research question and methodologies.

Literature reviews typically inform decisions around research directions. This is because literature reviews provide knowledge of unfamiliar research, redundant topics, new ideas, uncertainties and limitations, and a larger context of the topic [8]. In essence, literature reviews aid research decision makers with preventing some biases from entering the R&D process. However, there can be an overwhelming number of research publications on any given topic [9]. As a result, the time commitment for performing a literature review can be long, and in some cases, infeasible. The average time to review evidence-based medicine takes sixty-seven weeks [10], and the average review of psychological research takes six months [11]. Spending this much time on a literature review can cause the review to go out of date quickly because new research will be published as the review is undertaken [10]. **A solution to this problem is to use a sample of the research** publications to perform an inferential analysis to generalize the population of research being reviewed. Estimating a population using a sample requires the correct number of observations to avoid false negatives and false positives [12], but sample sizes are always smaller in quantity than the population size they represent. An additional benefit of inferential analysis is that hypothesis testing can be used to identify possible influences on the research trend [13]. Therefore, by collecting a sample of research on a specified topic, and inferring the quantity of articles by an aggregation unit (i.e., time and/or geography), it is possible to provide researchers with a method to quickly identify research trends, and reduce added biases.

Consequently, a rapid literature review was designed and tested using a sample of research publications in the field of public health. The research was specific to childhood elevated blood lead levels (EBLL) throughout the U.S. and published between 1990 and 2019. Our interests are particular to the environmental determinants of EBLL. Considering that research is influenced by R&D decisions, a research trend should be identifiable in this large group of research publications. The research publications were aggregated by research geography and publication year. The observed quantities of research, written by geography and year, were used to calculate an average number of papers per year concerning specific geography (i.e., states) Thus, this structure of analysis facilitates an inferential spatiotemporal analysis where explanatory variables, that are aggregated by the same geographical unit (i.e., research location) and time (i.e., publication year), can be used to describe the production of research over space and time. Currently, there are no analyses of research trends using inferential analysis of research location data. In fact, research trend analyses commonly use publication location and exploratory analysis methods. Furthermore, spatiotemporal analysis is appropriate for public health research because the patterns of diseases through time and space have been regarded as highly important but infrequently attempted [14]. Including temporal measurements into an analysis of research locations provides a novel approach for understanding the spatiotemporal trends of research. Additionally, if R&D publications represent technological change, it is possible that using R&D location will provide a measurement of the *impact of discovered ideas* on innovation. The *impact of discovered ideas* is a non-pecuniary measure that might represent the most interesting aspect of R&D social return spillovers [15]. Yet, quantifying this value has traditionally been very difficult. Therefore, the purpose of this analysis is to introduce an inferential analytical framework for a rapid literature review to identify research trends over time and space in a body of research, using EBLL research as a case study. The output of this analysis is intended to inform childhood EBLL R&D decision makers, stakeholders, and policy makers about the trends in childhood EBLL R&D and its production over time and space.

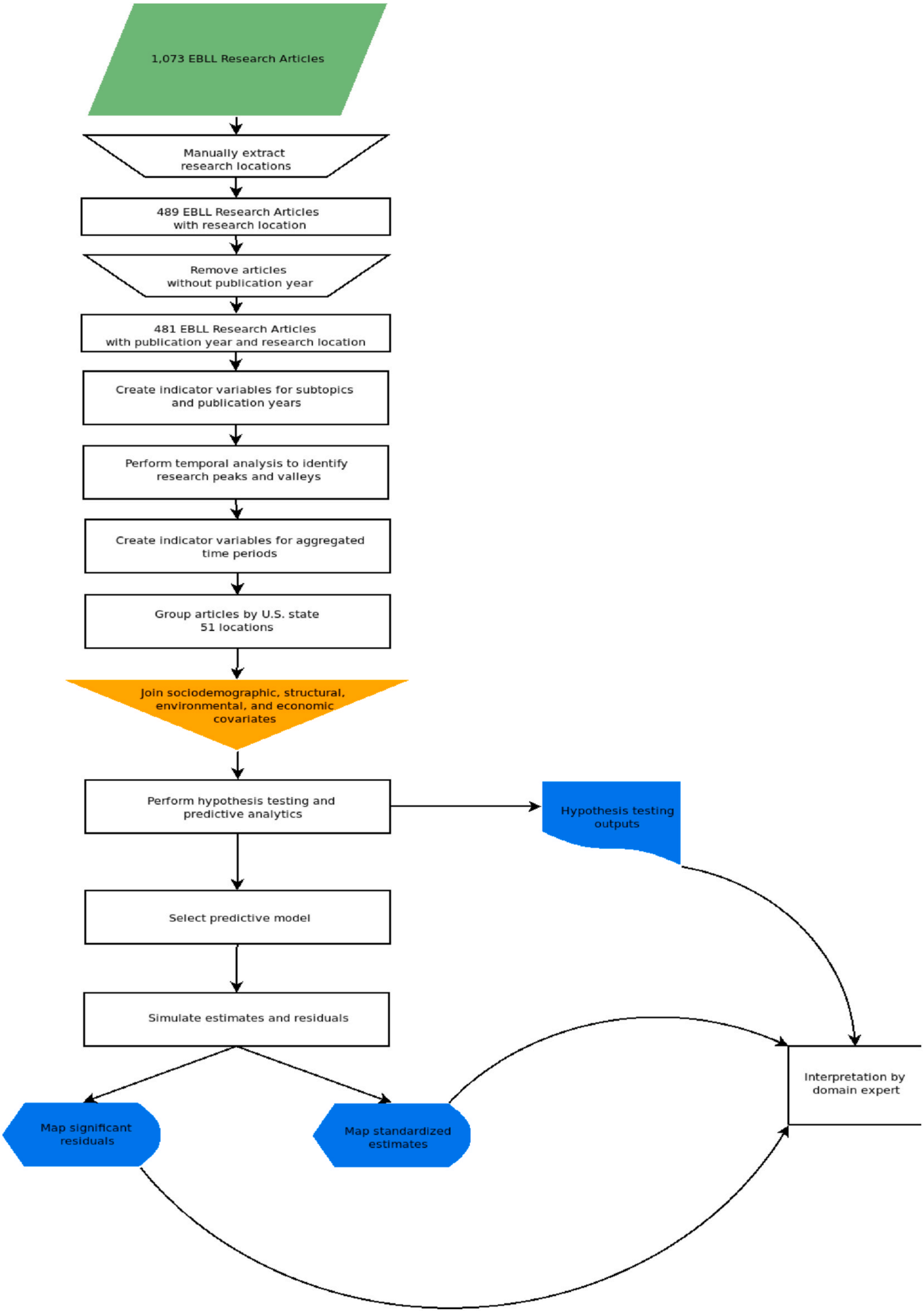

1,073 EBLL Research Articles
Manually extract research locations
489 EBLL Research Articles with research location
Remove articles without publication year
481 EBLL Research Articles with publication year and research location
Create indicator variables for subtopics and publication years
Perform temporal analysis to identify research peaks and valleys
Create indicator variables for aggregated time periods
Group articles by U.S. state 51 locations
Join sociodemographic, structural, environmental, and economic covariates
Perform hypothesis testing and predictive analytics
Hypothesis testing outputs
Select predictive model
Simulate estimates and residuals
Map significant residuals
Map standardized estimates
Interpretation by domain expert

Fig. 1. Flow chart depicting the process of data extraction, variable inclusion, hypothesis testing, estimation, visualization, and interpretation.

2. Methods

The purpose of this research is to define a framework that can be used to better understand a corpus of R&D publications using research locations, publication year, and article subtopics. The process for this framework included the following: 1) obtain a representative corpus of research publications for the childhood EBLL topic, 2) manually extract research locations and subtopics from article titles and abstracts, 3) extract the publication years from the article corpus, 4) calculate the annual frequency of articles per geographic unit, 5) calculate the frequency of articles per publication year, 6) create subtopic indicator variables and time-interval indicator variables, 7) obtain feature variables for analysis, 8) aggregate all data points (i.e., annual frequency of articles per geographic unit, subtopic and time-interval indicator variables, and feature variables) by geographic unit, 9) exploratory analysis, 10) statistical power analysis, 11) hypothesis testing, 12) estimation, 13) simulation, 14) visualization, 15) interpretation by domain experts. The combination of these steps form a methodology that can be used by R&D decision makers, stakeholders, or policy makers in any field that has a geographic dependency (i.e., public health, ecology, or economics)(Fig. 1).

The data used for this analysis was extracted from a body of public health research articles on the topic of childhood EBLL. The corpus of research publications was curated by the librarians at the U.S. Environmental Protection Agency as a representative list of available childhood EBLL research in the U.S. or Canada and written between 1990 and 2019. Three information content aggregators were used, by the librarians, to extract the body of research — PubMed, Web of Science, and ProQuest. Each of the information content aggregators were filtered for English language, scholarly articles and peer-reviewed articles. The search terms used for PubMed included: “child” or “child, preschool” or “infant” or “infant, newborn” or “adolescent” and “lead/blood.” The search terms used for Web of Science included: “blood lead” or “lead level*” or “lead exposure” or “exposure to lead” or “Pb exposure” or “Pb level*” or “environmental lead” and “children” or “pediatric.” And the search terms used for ProQuest included: “blood lead” or “lead levels” or “lead exposure” or “exposure to lead” or “Pb exposure” or “Pb levels” or “environmental lead” and “child*” or “pediatric.” The output from these three content aggregators, on the topic of childhood EBLL, in the U.S. and Canada, and between 1990 and 2019 was 1073 scientific journal articles (i.e., aggregated and de-duplicated by the U.S. EPA librarians). These articles were used to extract data for this analysis.

The first step for data extraction was to identify the geographic aggregation unit that would be used for the analysis. The criteria for selecting a geographic aggregation unit included 1) a unit that could easily be extracted from all the journal articles, and 2) a unit that is commonly used to aggregate other data. The U.S. state level geographic unit was selected. The decision to use this aggregation unit excluded Canada from the analysis. Next, the geography data was manually extracted from the 1073 childhood EBLL research articles. The articles’ titles and abstracts were manually reviewed, and the U.S. state where the research was conducted was extracted and documented; this is referred to as the “study location.” Study location does not necessarily align with the location of the author’s affiliation. Articles that did not have a study location were excluded.

In addition, article subtopics were manually extracted and identified as either “environmental” or “non-environmental” studies. Environmental studies were divided into subcategories: “indoor”, “outdoor”, or “both”. “Indoor” refers to potential indoor Pb sources and/or measurements inside a residence or structure, such as paint or dust concentrations. “Outdoor” refers to potential outdoor Pb sources and/or measurements, such as ambient air or soil measures. And “both” refers to studies that included both “indoor” and “outdoor” measurement and/or modeling components. Articles classified as “non-environmental” typically focused on blood Pb and health outcomes without any consideration of environmental exposure sources. This process was conducted by two separate researchers, independent of one another, and then validated for quality assurance and control. Validations were conducted by the researcher that did not decide on subtopics or research locations. All disputes were addressed through discussion between both researchers.

The articles were grouped by study location (i.e., U.S. state) and the frequencies of articles per U.S. state were calculated. The article frequencies were divided by the entire time-period (i.e., 29 years) to calculate the average number of articles written per year for each U.S. state. The average number of articles published each year, by research location, constituted the observations for the analysis. The subtopics (i.e., non-environmental, environmental, indoor, outdoor, and both) were also grouped by geography. The subtopic frequencies were then derived into binary indicator variables. This was done by replacing subtopic frequency values of 0 with a 0 and subtopic frequency values greater than 0 equal to 1 (e.g., a state with 0 articles written about a subtopic would be recorded as 0 while a state with 1 or more articles written about a subtopic would be recorded as 1). These variables were encoded as categorical values.

The result of the data extraction process was a dataset containing observations of the annual average of all scholarly articles on childhood EBLL, per U.S. state and the District of Columbia, and published between 1990 and 2019. Included with the observations were five indicator variables that represent whether the subtopics (i.e., non-environmental, environmental, indoor, outdoor, and both) were included in a published article in each U.S. state and the District of Columbia. Furthermore, time series analysis was achieved by grouping all the articles by publication year. The data was evaluated using a time series plot resulting in three distinct periods (1990–1995, 1996–2015, and 2016–2019) and confirmed using Chow tests. Thus, three additional indicator variables were derived to represent the time-period published. The Chow tests were calculated using linear regression of the publication year indicator variables. Consequently, the revised output dataset included a column of observed article yearly averages by research location and eight indicator variables (i.e., five subtopics and three time-periods).

In addition to extracting and grouping data from the research articles, independent variables were identified and obtained at the U. S. state level. Sociodemographic variables and structure ages have previously been correlated with EBLL [16,17,18,19]. Therefore, sociodemographic and structure age data were collected as potential explanatory variables. Additionally, economic and environmental variables were obtained, as these may also influence research production on childhood EBLL. All independent variables were collected for the timespan between 1990 and 2019, and the mean values per U.S. state were used. In some cases, the timespan was averaged by the available dates within the 1990 and 2019 time span (i.e., mean of data collected every three years or mean of data collected every decade). The primary data sources included the U.S. Census Bureau, the U.S. Bureau of Economic Analysis, the U.S. Bureau of Labor and Statistics, the U.S Environmental Protection Agency, the U.S. Department of Agriculture, the U.S. Geological Survey, and the National Science Foundation. Each independent variable was downloaded using either the governing body's data portal or an application portal interface (API). All available data for the time-period between 1990 and 2019 was downloaded and averaged to create the independent variable dataset. **This resulted in 89 independent variables (Table S-1)**. For the sociodemographic variables that were reported as total units, the per capita value was calculated to account for the total population. Also, independent variables with 5% or fewer missing data were imputed using simulations or mean values from the surrounding U.S. states. **If more than 5% of the data was missing the variable was removed from the analysis.** Imputation was used for social and economic factors and neighborhood averaging was used for physical factors. For example, a simulated imputation procedure was completed for Wyoming's "Management Earnings" variable, and the average values were calculated using Maryland's and Virginia's values for the District of Columbia's "Soil Pb Mean," and "Soil Pb Median" variables. It was done this way because Maryland and Virginia are the two U.S. states surrounding the District of Columbia. In this case we made the assumption that near things are more alike than far things [20].

Next, education and age of structure data were separately aggregated into derived variables. Sociodemographic factors, including maternal and paternal education level, may be associated with higher EBLL for children living in the home, therefore the variables representing the population based on education level were combined into two separate, continuous variables that measure the population of: 1) "High School Education and Lower" and 2) "Some College Education and Higher." Furthermore, the variables for the quantity of structures based on their year built were aggregated. **The aggregation thresholds were based on the dates Pb-based paints and leaded gasoline were historically used.** The variables that represent the quantity of structures built between 2005 or later, 2000-04, 1990-99, 1980-89, 1970-79, 1960-69, 1950-59, 1940-49, and 1939 or older, were combined into four derived variables: 1) "Structures Built After 1970"; 2) "Structures Built Before 1970"; 3) "Structures Built Before 1960"; and 4) "Structures Built Before 1950." Lastly, all independent variables were standardized to account for the vastly different scales between them. These new explanatory variables were merged with the observed data and indicator dataset to create a dataset that combined all observations (i.e., average number of articles written per year per U.S. state), all indicator variables (i.e., years before 1996, years after 2015, years between 1995 and 2015, non-environmental, environmental, indoor, outdoor, both), and 95 standardized explanatory variables (i.e., 89 independent variables plus two education variables and four housing age variables). The 51 rows of this data set represent the 48 contiguous U.S. states, Alaska, Hawaii, and the District of Columbia.

Before estimating the population of research papers, exploratory analysis of the observed data was conducted. This included summary statistics, bar charts, and maps. Additionally, power analysis was conducted to estimate the minimum sample size and statistical sensitivities, for this analysis. Then, hypothesis testing was performed to determine whether the independent variables influenced the generalized population compared with random chance [8,12]. The Spearman correlation was calculated between observations and explanatory variables for each U.S. state. A univariate general linear model for regression analysis was used for hypothesis testing. To account for the influence from outliers, a logarithmic transformation was applied to the sample observations [12]. The Benjamini & Hochberg false discovery rate was used to adjust the p-values. Variables that significantly occurred for reasons other than random chance were selected and added to a list of significant variables. This list represented variables that suggest an effect on the production of these research publications. Together, the correlation analyses and the list of suggested effects were used to guide the predictive modeling approach.

To begin the predictive modeling analysis, the list of significant variables was reduced by excluding variables that were highly correlated with each other. The recommended threshold for high correlation used was $r \geq 0.90$ or $r \leq -0.90$ [12]. Therefore variables that were highly correlated were separated into two datasets, and the correlated variables were later compared during model specification. For example, if a model specified "Structures Built Before 1950" as a covariate, other models were compared using the highly correlated structure age variables (i.e., "Structures Built Before 1960" or "Structures Built Before 1970"). Variable selection was automated using an exhaustive search for the best covariate subset using a branch-and-bound algorithm [21]. The model specification procedure was completed with the following steps. First, a model was specified using no fixed effects. Next, four models were specified using each of the derived subtopic variables as fixed effects. Three additional models were specified using the three derived time-period variables as fixed effects. Lastly, all specified models were analyzed using an analysis of variance (i.e., ANOVA) table. Within each model specification step, models were compared with the variables that were highly correlated with the explanatory variable identified by the exhaustive search algorithm, polynomial models, and interaction models. These comparisons allowed us to verify which of the related variables fit best and avoid multicollinearity. Validation was performed using Maximum Log Likelihood estimates (LL), Bayesian Information Criterion (BIC), Cook's distance, and cross validation. Also, the specified models (i.e., one model with no fixed effects, four models with subtopic fixed effects, three models with time-period fixed effects, and one model with one subtopic fixed effect and one time-period fixed effect) were validated among each other using an ANOVA table. This aided the decision to choose a single predictive model based on *a priori* knowledge, F-values, BIC, and LL. The final model selected was validated for independence using Moran's I. The regression trends were simulated to account for stochasticity, and the results were standardized. Finally, the article production per state was mapped in bins based on their standard deviation's magnitude and direction from the mean of the estimated population of EBLL R&D production.

All analyses were produced using R 4.1.2 (2021-11-01) statistical computing and graphics language. The integrated development environment (IDE) used was Rstudio 2022.12.0. R Markdown 2.20 was used to format code chunks and produce dynamic code and commentary documents. Power analysis was calculated using the *pwr* library. The *leaps* library was used for the branch-and-bound algorithm to identify general linear models with the lowest BIC score.

3. Results

The literature search returned 1073 scientific journal articles on childhood EBLLs. After extracting research locations, there were 489 articles (i.e., 592 articles were excluded because there was no research location associated with them). Eight articles did not have a publication year recorded, which were removed, leaving 481 articles that comprised the observed or sample dataset (Fig. 2). The observed range of the annual average of research articles by state was 0–1.7. The observed mean annual average of articles was 0.3. Seven U.S. states had no publications: Connecticut, Delaware, Oregon, Tennessee, Wyoming, South Dakota, and Hawaii (Fig. 2). The states with the highest average number of articles written per year were New York, Michigan, California, Louisiana, and Massachusetts (Fig. 2). The proportion of states with a research location and focused on the environmental subtopic was 65%. Among these, 35% focused on the indoor environment, 57% focused on the outdoor environment, and 31% focused on both the indoor and outdoor environment.

Trends over time were plotted and analyzed. These plots showed two peaks of research on childhood EBLLs in 1995 and 2016 (Fig. 3). Chow tests were used to confirm these results (p-values <0.001). As there is typically a lag between research being conducted and a journal article being produced, these peaks likely reflect the change in the CDC's level of concern for children's blood lead levels to <10 μg/dL in 1990,[25] and the Flint water crisis in 2014.[26] This coincides with an average 3 year lag found in production and innovation [15]. The proportion of states and the District of Columbia with articles published between 1990 and 1995, 1996–2015, and 2016–2019 was 41%, 84%, and 65% respectively.

Thirty-eight explanatory variables significantly varied from random chance. These were grouped into their broader categories: temporal, sociodemographic, education, structure age, environmental, and economic variables (Table 1). All three temporal indicator variables, eight sociodemographic variables, ten educational variables, eleven structure age variables, two environmental variables, and four economic variables suggest a significant non-random effect on the yearly average production of childhood EBLL articles per U. S. state. For the sociodemographic variables, total population, U.S. citizenship, and poverty measures were variables with a suggestive effect on EBLL article production. Also, it is noteworthy that two environmental variables (i.e., atmospheric Pb in the PM_{10} size range and the number of National Priority Sites [NPL]) were statistically significant. The results suggest that states with higher populations, lower income households, older structures, more environmental Pb contamination, and higher investment in research are more likely to produce research on this topic. Also, lower education can be a contributing factor for higher EBLL, but higher education is more likely to produce research publications. The effect of education and EBLL is typically studied at the individual or community level, but it is possible that this trend is consistent across geographic scales, and reflected in this analysis at the U.S. state resolution. If these findings are correct, it suggests that at the state level, both low education and high education have a positive effect on the average number of papers written on childhood EBLL each year. This demonstrates an important relationship for future research. The only negative relationship observed was the "Median Year Structures Built' variable (i.e., coefficient of −0.074). This suggests that states with newer construction (i.e., larger values indicate newer construction) have less research production for this topic compared to states

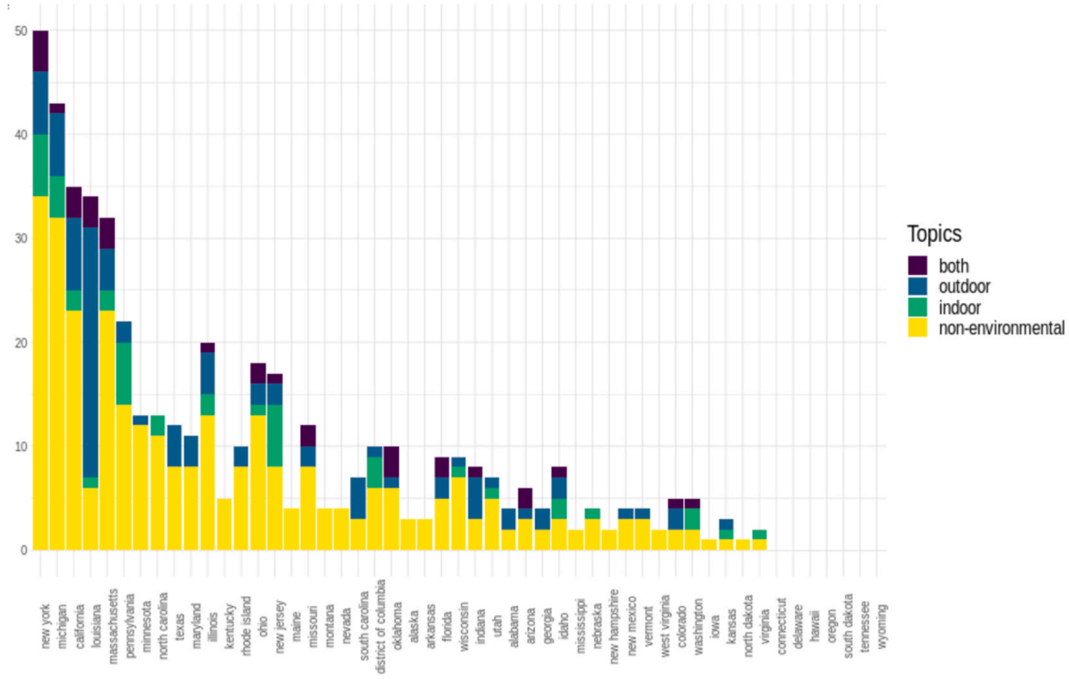


Fig. 2. Barchart of observed quantities of childhood elevated blood lead level (EBLL) research articles written between 1990 and 2019 at research location (U.S. state). Color segments represent the quantity of subtopics the overall article frequencies are made up of.

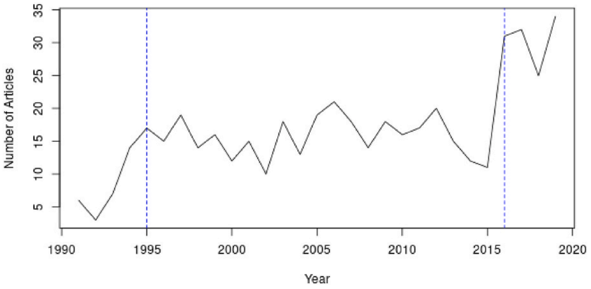


Fig. 3. Time trend of observed quantities of elevated blood lead level (EBLL) research articles from 1990 to 2019. Dashed blue lines depict two peaks of article production in 1995 and 2016.

with older buildings. This might relate to historic city centers need to study this topic due to industrial use of lead and older infrastructure that relied on lead usage. The findings from the two environmental explanatory variables suggest that researchers are studying EBLLs in areas with higher Pb concentrations in the air or they are studying areas with more U.S. EPA Superfund sites. This might signify areas with higher quality blood lead data provided by states and state organizations promoting this research. The Comprehensive Environmental Response, Compensation and Liability Act (CERCLA) is informally called Superfund, and it is a law that allows the U.S. EPA to clean up contaminated sites at the expense of the responsible parties [22]. These sites are not specifically contaminated by Pb, but Pb is a contaminant that is mitigated at many of these sites. Also, these sites could act as surrogates for general contamination and communities at greater risk of chemical exposure. Hence, the outcome that these sites influence EBLL research production has a reasonable explanation. Likewise, higher Pb concentrations in the air might translate to a stronger need to study the subject, but it does not explain why the other Pb environmental variables did not show a significant effect. This too represents an area for further study. The strongest suggested effect on EBLL research production, outside of the time-interval variables, was from the derived variable "Structures built Before 1950" (i.e., coefficient of 0.137), and it coincides with the results observed from the median structure age, and from the *a priori* knowledge that older structures can be a significant exposure source for EBLL [16,17,18]. Again, this might represent historic city centers' influence on this research.

Nine predictive models were used to generalize the population of the average number of articles written in each U.S. state on childhood EBLLs (Table 2). In all nine models, the variable "Structures Built Before 1950" was used as a covariate. The exhaustive search algorithm selected it as a covariate with the lowest BIC score for all the models analyzed. An additional group of models that used the highly correlated variables were calculated and compared with the first model that used the covariate "Structures Built Before 1950." The highly correlated variables compared were "Structures Built Before 1960" and "Structures Built Before 1970." These three models were compared with the lower order linear model and the higher order quadratic model using the covariate "Structures Built Before 1950" with an ANOVA table. The quadratic model that used "Structures Built Before 1950" was the only model to significantly vary from random chance (i.e., p-value <0.05) compared with the lower order models. Also, the model residuals were normally distributed (S-3). Therefore, the quadratic model using "Structures Built Before 1950" was selected as a baseline predictive model (Fig. 4). Similarly, all three time-period models were compared, and no significant difference was identified. One additional model was included that contained fixed effects for subtopics and time-period. It included the environmental subtopics and the 1990-95 period because the environmental subtopic was the biggest interest in this analysis, as well as how the research was influenced by change (i.e., government regulation and Pb contamination events). Each model was compared to the baseline model. The nine models were analyzed using an ANOVA table (S-2) and the selected model was the model that included the environmental subtopic indicator variable, time-period 1990-95 indicator variable, and "Structures Built Before 1950" covariate. The selected model was chosen because it significantly varies from random chance (i.e., p-value <0.05), has one of the lowest BIC scores equal to -44.00, the highest Log Likelihood value of 31.94, it represents the two variables of interest for the analysis (i.e., publication time-interval and subtopic), and the model residuals are normally distributed (S-4).

The selected model estimated an average of 0.45 articles written per year and a range of 0.21–1.16 articles per year, by research location. The model included three variable combinations (i.e., covariate plus two binary indicator variables) that significantly varied from random chance. These were: time-period 1990–1995 and non-environmental subtopic; time-period other than 1990–1995 and environmental subtopic; and time-period other than 1990–1995 and non-environmental subtopic. Only one state fell into the category of time-period 1990-95 and non-environmental subtopic. This variable combination did not significantly vary from random chance. The selected model's significant regression lines were plotted and analyzed (Fig. 5). It was observed that all U.S. states that contained no EBLL articles on environmental subtopics were states whose quantity of structures built before 1950 are less than the mean quantity of structures built before 1950 (Fig. 5-A). In other words, this model suggests that states with a quantity of structures built before 1950 that are below the average tend to not focus on environmental research for this topic. Also, the regression lines suggest that the U.S. states having articles produced between 1990 and 95, on environmental subtopics, tend to produce the most articles per year on childhood EBLL (Fig. 5-C). This may be explained by a given researcher or university having a long-standing study and/or community relationship in a particular area. Or perhaps this is because of a persistent contamination issue. In contrast, states with articles published within 1990-95 and not focused on environmental subtopics tend to have the least articles per year written on childhood EBLL (Fig. 5-B). This contrast, in time periods and subtopics, suggests that either states that were the focus of EBLL research before the CDC

Table 1
Table of hypothesis tests with significant occurrence compared with random chance.

Independent Variable	Coefficient	Z - Value	P - Score	Adj. P - Score
Wrote an EBLL article within 1991 and 1995	0.19	4.47	< 0.0001	0.0002
Wrote an EBLL article within 1996 and 2014	0.24	3.93	0.0003	0.0008
Wrote an EBLL article within 2015 and 2019	0.2	4.5	0.0001	0.0001
Total Population	0.11	5.33	< 0.0001	< 0.0001
US Citizenship (Born in US)	0.11	5.53	< 0.0001	< 0.0001
Below Poverty within 12 months	0.1	4.92	< 0.0001	< 0.0001
Gini Index for Income Inequality	0.08	3.35	0.0016	0.0040
Total Housing Units	0.11	5.53	< 0.0001	< 0.0001
Vacant Housing Units	0.1	4.41	0.0001	0.0002
Owner Occupied Housing Units	0.11	5.58	< 0.0001	< 0.0001
Total Residents per Housing Unit	0.11	5.36	< 0.0001	< 0.0001
Federal R&D Revenue	0.08	3.48	0.0011	0.0030
State R&D Revenue	0.09	4.38	0.0001	0.0002
State R&D Expenditure	0.1	4.73	0.0000	0.0001
State Venture Capital per GDP	0.08	3.46	0.0011	0.0031
Less Than High School Degree	0.1	4.49	< 0.0001	< 0.0001
High School Degree	0.11	5.81	< 0.0001	< 0.0001
Some College Education	0.1	5.01	< 0.0001	< 0.0001
Bachelor's Degree	0.11	5.43	< 0.0001	< 0.0001
Graduate Degree	0.12	6.04	< 0.0001	< 0.0001
High School Degree or less	0.11	5.4	< 0.0001	< 0.0001
Some College or more	0.11	5.39	< 0.0001	< 0.0001
Count of S&E Articles Published	0.13	7.41	< 0.0001	< 0.0001

Bachelor's Degrees in Workforce	0.12	5.93	< 0.0001	< 0.0001
Graduate Students per 1,000	0.13	7.1	< 0.0001	< 0.0001
Structure Built 1980-1989	0.08	3.28	0.0019	0.0047
Structure Built 1970-1979	0.09	3.89	0.0003	0.0009
Structure Built 1960-1969	0.11	5.11	< 0.0001	< 0.0001
Structure Built 1950-1959	0.12	6.64	< 0.0001	< 0.0001
Structure Built 1940-1949	0.13	7.36	< 0.0001	< 0.0001
Structure Built 1939 or Earlier	0.14	8.03	< 0.0001	< 0.0001
Median Year Structure Built	-0.07	-3.21	0.0024	0.0056
Structure Built After 1970	0.08	3.29	0.0019	0.0047
Structure Built Before 1970	0.13	7.27	< 0.0001	< 0.0001
Structure Built Before 1960	0.13	7.94	< 0.0001	< 0.0001
Structure Built Before 1950	0.14	8.29	< 0.0001	< 0.0001
Atmospheric Pb (PM10)	0.08	3.43	0.0012	0.0033
National Priority List Sites	0.12	6.81	< 0.0001	< 0.0001

Table 2
Table of models specified and compared by cross validation (CV), BIC, and LL.

Model	RMSE – CV	COD (R^2) – CV	BIC	LL
ŷ ~ Structures Built Before 1950 + (Structures Built Before 1950)	0.16	0.59	−33.66	24.79
ŷ ~ Structures Built Before 1950 + Environment Subtopic	0.14	0.68	−45.48	30.68
ŷ ~ Structures Built Before 1950 + Indoor Environment Subtopic	0.15	0.63	−40.96	28.45
ŷ ~ Structures Built Before 1950 + Outdoor Environment Subtopic	0.14	0.69	−47.52	31.73
ŷ ~ Structures Built Before 1950 + Both Indoor and Outdoor Subtopic	0.14	0.66	−44.17	30.10
ŷ ~ Structures Built Before 1950 + 1990–1995	0.15	0.64	−40.51	28.22
ŷ ~ Structures Built Before 1950 + 1996–2014	0.14	0.67	−43.99	29.93
ŷ ~ Structures Built Before 1950 + 2015–2019	0.14	0.65	−42.21	29.05
ŷ ~ Structures Built Before 1950 + Environment Subtopic + 1990–1995	0.14	0.69	−44.00	31.94

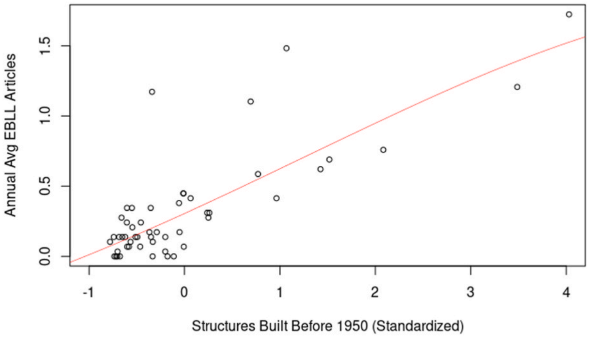
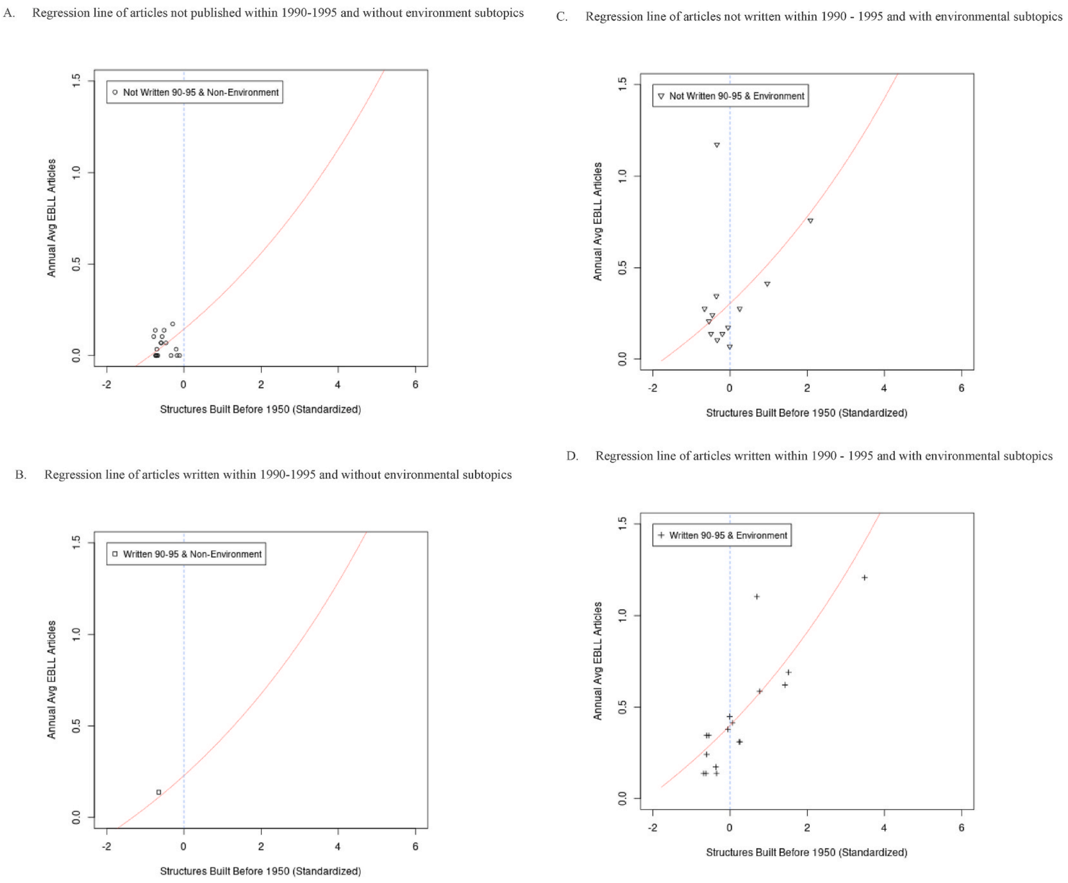


Fig. 4. Quadratic linear model using dependent variable (Annual Average of EBLL Articles per U.S. state) regressed to independent covariate 'Structures Built Before 1950', and used as baseline model.


A. Regression line of articles not published within 1990-1995 and without environment subtopics

C. Regression line of articles not written within 1990 - 1995 and with environmental subtopics

B. Regression line of articles written within 1990-1995 and without environmental subtopics

D. Regression line of articles written within 1990 - 1995 and with environmental subtopics

Not Written 90-95 & Non-Environment
Not Written 90-95 & Environment
Written 90-95 & Non-Environment
Written 90-95 & Environment
Annual Avg EBLL Articles
Structures Built Before 1950 (Standardized)
0.0
0.5
1.0
1.5
-2
0
2
4
6

Fig. 5. Selected fixed effects model using dependent variable (Annual Average of EBLL Articles per U.S. state) regressed to independent covariate 'Quantity of Structures Built Before 1950' and fixed effects 'Environmental' subtopics and time-period '1990–1996'. The vertical blue line represents the mean value of structures built before 1950 across all U.S. states analyzed. A. Regression line of articles not published within 1990–1995 and without environment subtopics B. Regression line of articles written within 1990–1995 and without environmental subtopics. C. Regression line of articles not written within 1990–1995 and with environmental subtopics. D. Regression line of articles written within 1990–1995 and with environmental subtopics.

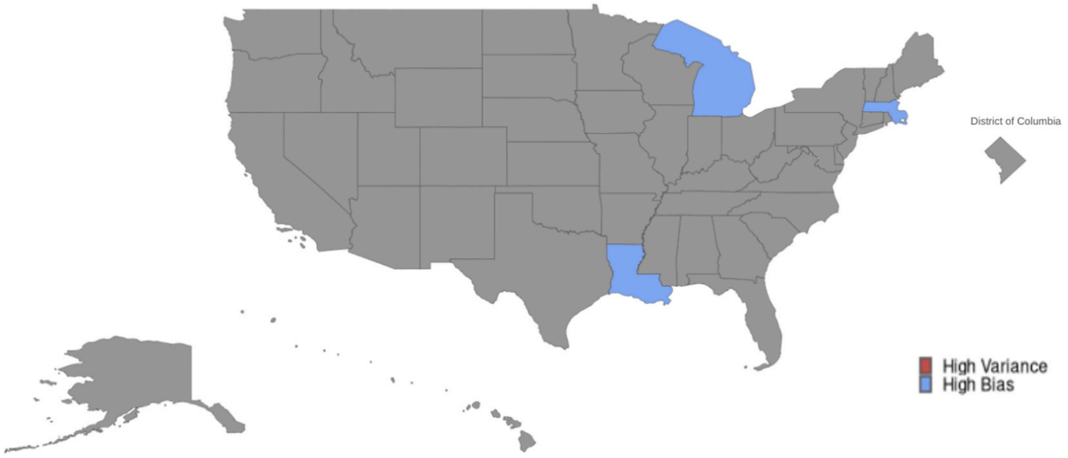


Fig. 6. Map of U.S. states where the simulated model contained significantly high or low bias or variance.

changed the BLL threshold have a propensity to research environmental conditions, or states that have a propensity to research environmental conditions tend to study this topic of EBLL historically.

The results of the selected model were simulated and plotted on maps to visualize the bias-variance trade off and geographic distribution of the research production. First, the simulated residuals were standardized and plotted to identify significantly high bias or variance (Fig. 6). The map showed Louisiana, Michigan, and Massachusetts to be statistically biased by the model. The variance was not statistically significant. This signifies a useful and parsimonious model. Next, the simulated estimates were standardized and plotted to view the geographic distribution of childhood EBLL research (Fig. 7). According to this map, California and New York tend to produce the highest amount of research on this topic each year, based on the time-period, subtopic, and structure age. States that produce this research within one standard deviation above the mean estimates include Illinois, Michigan, Ohio, Pennsylvania, and New Jersey. Lastly, states below the mean estimates include Alaska, Hawaii, Nevada, Wyoming, North Dakota, South Dakota, Mississippi, Delaware, and New Hampshire.

4. Discussion

This methodology demonstrated an improvement from previous research on research trends. Particularly, this research follows Millard et al.'s research on research disparities. As discussed previously, Millard et al. identified geographic and taxonomic disparities in animal pollinator literature [5]. They attribute both the geographic and taxonomic disparities to data availability and representation bias. Further inspection revealed that the geographic disparities appear to follow economic trends when the top producers of the animal pollinator literature are compared with national GDP for the top producing countries.[27[Considering this observation, it became clear that an observational, inferential analysis of R&D research locations might reveal more nuanced trends in the data. Furthermore, Millard et al. produced their results using automatically extracted data. We felt that it was important to manually extract the data for this analysis to ensure data representativeness. Using misrepresentative data would result in an incorrect estimate of the population of these EBLL R&D publications. By using representative data, we have a baseline analysis for future research on performing this analysis with automatically extracted data.

The resulting analysis identified social events that affected production by time-intervals (i.e., the CDC blood lead level threshold change in 1990 and the Flint Water Crisis in 2014), and it identified potential effects on why research is conducted in certain locations (i.e., significant hypothesis testing results). Additionally, it estimated the population of research, by research location and time-interval, and displayed these in maps. The outputs provide a novel approach to investigating R&D. This problem affects all researchers in varying fields, but it is limited to fields with research locations. This methodology is relevant to R&D decision makers, stakeholders, and policy makers because it can provide quantitative results about a geographically dependent research field across time and space. These quantitative results can help alleviate R&D bias by providing strong tools to support a more quantitative decision making process for R&D planning.

This case study is highly relevant to childhood EBLL researchers, stakeholders, and other EBLL decision makers. Interpreting these results requires domain knowledge in this field. An example of how this might help decision makers identify new research opportunities for childhood EBLL is by looking at the geographic disparities, and identifying where research is over-concentrated versus under-concentrated. New York and California, for instance, have a high degree of research on this topic. Considering that structure age has a strong effect on this research, and that both New York and California have historic cities that relied on lead infrastructure and industrial process, it might be of interest for research to identify similar states that have lower levels of research. However, researchers need to be aware of the potential reasons that could justify more research in New York and California, as demonstrated previously. Clearly, external factors outside of the suggested effects might exist when deciding upon research directions. These may include considerations such as funding sources, location accessibility, willingness of communities to participate, and state policies on pediatric

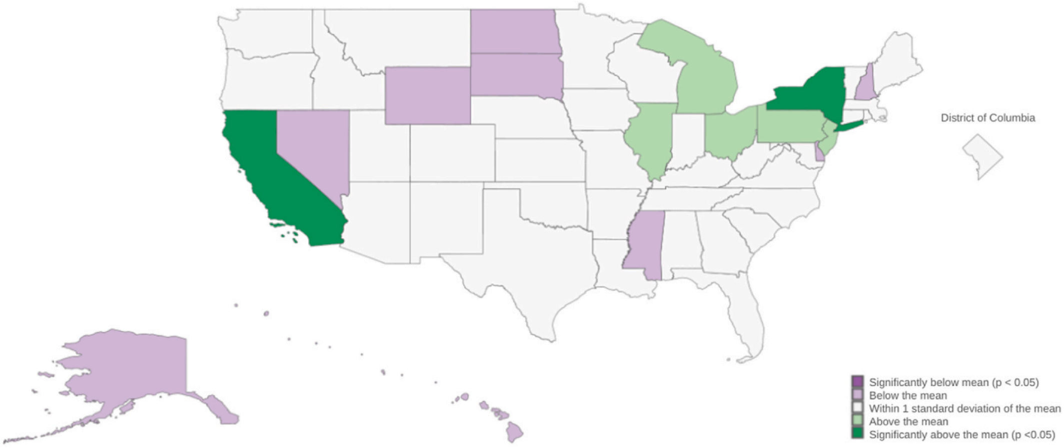

District of Columbia
Significantly below mean (p < 0.05)
Below the mean
Within 1 standard deviation of the mean
Above the mean
Significantly above the mean (p <0.05)

Fig. 7. Map of U.S. states using simulated model estimates to identify U.S. states average level of EBLL research productivity.

lead screening. Only a researcher familiar with these details will be able to use these results to make accurate decisions about whether the knowledge gaps identified follow reasonable research trends in the environmental EBLL field. But it must be considered that the most experienced EBLL researcher might equally uncover new locations to research EBLL as they might fall victim to confirmation bias or follow research trends because of convenience. In essence, these results generate questions that can help researchers diverge from unnecessary trends or conform to necessary ones. Examples of these questions are: 1) Why do U.S. states with newer housing stock tend to not publish research about environmental subtopics related to childhood EBLL? 2) Why do U.S. states with publications prior to 1996 and about environmental subtopics tend to be states that produce the most EBLL research? 3) Why is it that older structures, atmospheric Pb in the PM_{10} size range, or NPL sites influence this research? Hence, research decision makers can use these outputs to find new opportunities for researching this topic, or stick to those that are verified by this process, but domain expertise is necessary.

Another aspect of this research is that it potentially contributes to measuring the *impact of discovered ideas*, or the social return spillover from R&D. *Discovered ideas* represent the hardest measure of social return spillover [15]. It also represents the most informative. However, attempts to capture it have not been successful because of its non-pecuniary scale. One attempt to capture spillover used a spatial lag model with firm level expenditures, across the U.S., and at the U.S. state scale [23]. The methodology did a good job of following the recommendations by leading researchers, but it used financial data to model production. Financial data is a poor measure for social return [15], and therefore the research in that study did not do a good job of capturing the *discovered ideas* social return spillovers. Alternately, the method presented here captures a quantifiable, non-pecuniary metric of research by research location, which describes production and innovation as R&D quantity to represent technological change. Furthermore, this analysis provides temporal, sociodemographic, economic, environmental, and structural effects on these metrics. As a result, this analysis represents an analysis of *discovered ideas* social return spillover across the U.S., at the state scale. Hence, the suggested effects and population of estimates identify trends in the indirect effect of technological changes for this topic. As a result, this approach also affects the study of innovation and production.

Furthermore, there are many categories of bias that can influence research trends. It is beyond the scope of this analysis to identify which types of bias contribute to these trends, as the goal of this analysis was to identify the trends for interpretation by domain experts. However, it is prudent to mention that bias in research is always a possibility, and it can be intentional or unintentional [24]. In research, bias can influence multiple steps in the process, including data collection, data analysis, interpretation, and publication [24]. Our presented methodology may be biased in some decisions that were made. The search terms and engines clearly influenced the research articles that were returned. Due to the limitations of the content aggregators, there may have been a handful of articles missed during the literature search but given our objective to develop a methodology that uses a sample of these articles to predict a population, this is acceptable. It also means that it is crucial to define the population of papers up front that one is striving to generalize through careful consideration of search terms. The selection of the geographic unit at the U.S. state level could also introduce bias. Environmental EBLL research is typically conducted at a higher geographic resolution (i.e., the community scale), and so applying the methodology at the census block resolution might better suit the intended purpose. However, a finer resolution may not be easily extractable nor be consistent across studies (e.g., zip codes versus census tracts), severely limiting the ability to do the analysis. Next, selecting the subtopics is a subjective decision because the domain experts must choose, and there are often multiple ways to categorize research papers from a broad literature search. Additionally, the analysis results can be biased if the researchers applying the methodology only report significant values or unnecessarily remove outliers [24]. Interpretation is subjective too, as different researchers may reach varying conclusions based on their experiences and training. And finally, publication bias is another consideration [24]. **Publishers might be more inclined to publish articles with significant results,** providing greater numbers of papers with interesting, quantifiable results, when in fact, just as much can be learned from research with non-significant findings [24]. For example, in the environmental EBLL literature, states that may have conducted Pb studies but did not find any relationship nor publish on it, could be shown as underrepresented. This methodology was developed to improve R&D decision making time commitments while mitigating bias to the extent possible. Yet there is always the potential that bias can infiltrate the process, and it is crucial to recognize that this is not a foolproof methodology. Therefore, the best method for reducing bias is for researchers to be more introspective when planning research. Furthermore, causation cannot be stated from this analysis because this is an observational study and as a result, the hypothesis tests can only suggest an effect on article production that does not occur from random chance. Also, it is important to keep in mind that these results can only be interpreted for the U.S. state aggregation unit [20]. There are circumstances where understanding the U.S. state trends are necessary, such as decision making on a national scale. The U.S. state level analysis was justified for this analysis as both a proof of concept and a practical case study on a topic that has been extensively studied over the past 40 years.

Additionally, data representativeness is always a concern for statistical analysis. In this analysis, the literature search was performed by professional librarians that work for a federal research institution. They provided this corpus of publications with the intent of identifying all available research in the location and time period of interest. Understanding this is necessary for any researcher that might use this methodology because a literature search that does not represent the population will provide incorrect results, and an incorrect analysis. This is not unlike performing a literature review, where the literature search is expected to be a comprehensive representation of publications about a topic.

Another important limitation for this analysis is the feature variable data representation. In this analysis it was necessary to use the average values for the data over the time span of interest (i.e., 1990–2019). We assumed that the mean value of this data was representative of the variables of interest. This is also relevant for the data imputation process, which assumed that the mean value for all locations when viewing sociodemographic variables, and the mean value of the nearest locations when looking at environmental variables, represent the variable of interest. Ideally, all of the variables of interest would have data for each year in the time span, and it would be separated by time interval, but this was not possible with the available data sets.

5. Conclusion

The purpose of time trend analysis, hypothesis testing, and predictive modeling across research locations was to provide R&D decision makers with analytical tools they can use to assess existing literature as part of the research planning process, while mitigating bias. According to one of the largest global consulting firms, R&D decisions must be made quickly for competitiveness [25]. Yet, fast decisions tend to introduce human bias into the decision making process [26]. It is the slower, analytical decisions that tend to reduce human bias [26]. This methodology provides a framework for quickly influencing analytical decisions, and therefore allows decision making with less human error. In providing this framework, a novel approach for spatiotemporal analysis of R&D by research location was demonstrated. The result provided suggested effects and estimates of R&D production by research location, which might also capture the *impact of discovered ideas*. As a result, the methodology that was demonstrated has strong implications in understanding research trends, what factors affect research trends, the literature review process, and measures of production and innovation.

The demonstrated methodology used a standard literature search with more than 1000 research articles and returned a table of variables that suggested significant relationships with the number of research articles on this topic per year, and in each research location. It also generalized the research from the sample of articles to **an infinite population of articles** written each year, and in each research location. These estimated values were used to identify research locations, at the U.S. state resolution, where over-versus under-represented research is being conducted. By combining these results, we identified the suggested influences on this research and location trends for conducting this research. The suggested influences fell into the temporal, sociodemographic, economic, education, structure age, and environmental factors categories. Structure age appeared to be the best estimator. The estimated annual research papers were highest in New York and California and lowest in Alaska, Hawaii, Nevada, Wyoming, North Dakota, South Dakota, Mississippi, Delaware, and New Hampshire. Additionally, residual analysis showed that the model underestimates articles written in Louisiana, Michigan, and Massachusetts. These estimates and underestimates identify gaps in geography for the body of environmental EBLL research. Furthermore, these results raise questions as to why Louisiana, Michigan, and Massachusetts produce more articles than the estimated trend.

The outputs of this analysis (i.e., suggested effects and estimated annual production) offers childhood EBLL researcher decision makers a set of quantitative tools to inform the R&D decision making process. Interpretations of this data must be made by domain experts. However, anyone can understand the maps that display New York and California as the significantly largest producers of childhood EBLL research, by research locations, and compare these with other lower producer states. This clearly demonstrates disparities in the research by research location. As stated before, these disparities should be assessed by domain experts. Additionally, these disparities can possibly be interpreted as *impact of discovered ideas* social return spillovers. Due to the technological change that R&D represents, and its use for measuring innovation [15], with the non-pecuniary measure this analysis uses, this analysis likely captures the indirect social return of childhood EBLL research. In response to this, this research is relevant to understanding production and innovation. However, more research should be performed to assess its potential. Overall, the biggest benefit from this research is that it provides researchers, research decision makers, stakeholders, and policy makers with more nuanced questions about a research topic of interest. These questions can be further assessed to benefit R&D quality, and mitigate bias in the decision making process.

The most time-consuming part of this analysis was the data extraction process. To manage this time commitment, the intent is to continue developing this methodology with text-mining applications to improve the speed of extracting data. Incorporating text-mining applications that can extract publication dates, research locations, and subtopics will greatly reduce the time requirements. These time requirements have already been reduced by using samples of research to generalize an infinite population of research. And, although the trend in research is to speed up the decision-making process, it is of the utmost importance that judgment errors are prevented in the process. It has been shown that fast decision making incorporates more human error. Using statistical analysis is a solution for this problem because it uses less data to make accurate decisions that can be validated. This methodological framework shows that statistical inference reduces the time needed for reviewing research papers using publication year, research locations, and subtopics. Pairing a text-mining computer application for data extraction with this inferential methodology will make this process even faster.

Data availability

The data that support the findings of this study are openly available in the repository named ‘data’ at https://github.com/nickgrok/data.

CRediT authorship contribution statement

Nicholas Grokhowsky: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This manuscript was prepared using data that was extracted from a literature search conducted by the U.S. Environmental Protection Agency's librarians and does not necessarily reflect the opinions of the U.S. Environmental Protection Agency or its staff. I appreciate the contribution by Lindsay Stanek of the U.S. EPA's Office of Research and Development, Center for Public Health and Environmental Assessment. Additionally, I thank Rogelio Tornero-Velez of the U.S. EPA Office of Research and Development, Center for Computational Toxicology and Exposure.

Appendix

Table S-1
List of variable downloaded with their source

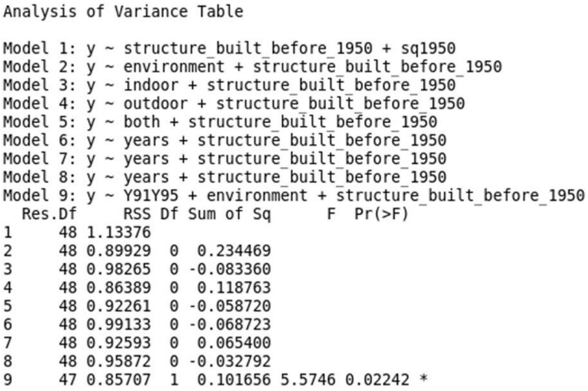
Variable	Source
Total Population	US Census Bureau
Male Population	US Census Bureau
Female Population	US Census Bureau
Age Under 18 Population	US Census Bureau
White Population	US Census Bureau
Black Population	US Census Bureau
American Indian Population	US Census Bureau
Asian Population	US Census Bureau
Pacific Islander Population	US Census Bureau
Other Race Population	US Census Bureau
Multiple Race Population	US Census Bureau
Hispanic Population	US Census Bureau
Median Age Population	US Census Bureau
US Citizen Population	US Census Bureau
Below Poverty Population (Last 12 Months)	US Census Bureau
Median Household Income (Last 12 Months)	US Census Bureau
Gini Index for Income Inequality	US Census Bureau
Mining Earnings	US Census Bureau
Agriculture Earnings	US Census Bureau
Construction Earnings	US Census Bureau
Manufacturing Earnings	US Census Bureau
Wholesale Earnings	US Census Bureau
Retail Earnings	US Census Bureau
Transportation & Warehouse Earnings	US Census Bureau
Utilities Earnings	US Census Bureau
Information Technology Earnings	US Census Bureau
Finance & Insurance Earnings	US Census Bureau
Science and Technology Earnings	US Census Bureau
Management Earnings	US Census Bureau
Administrative Support and Waste Management Earnings	US Census Bureau
Education Earnings	US Census Bureau
Health & Social Care Earnings	US Census Bureau
Art, Entertainment, & Recreation Earnings	US Census Bureau
Other Services Earnings	US Census Bureau
Public Administration Earnings	US Census Bureau
Total Housing Units	US Census Bureau
Total Vacant Housing Units	US Census Bureau
Total Owner Occupied Housing Units	US Census Bureau
Median Home Value	US Census Bureau
Median Real Estate Taxes Paid	US Census Bureau
No. Structures Built After 2004	US Census Bureau
No. Structures Built 2000–2004	US Census Bureau
No. Structures Built 1990–1999	US Census Bureau
No. Structures Built 1980–1989	US Census Bureau
No. Structures Built 1970–1979	US Census Bureau
No. Structures Built 1960–1969	US Census Bureau
No. Structures Built 1950–1959	US Census Bureau
No. Structures Built 1930–1949	US Census Bureau
No. Structures Built 1939 or Earlier	US Census Bureau
Median Year Structure Built	US Census Bureau
Population Education Less Than High School	US Census Bureau
Population with High School Education	US Census Bureau
Population with Some College	US Census Bureau
Population with Bachelor's Degree	US Census Bureau

(*continued on next page*)

Table S-1 (*continued*)

Population with Graduate School Education	US Census Bureau
Total People Per Household	US Census Bureau
Unemployment Rate	Bureau of Labor and Statistics
Air Quality Pb in Particulate Matter (1.0)	US Environmental Protection Agency
Air Quality Pb in Particulate Matter (2.5)	US Environmental Protection Agency
Air Quality NO_2	US Environmental Protection Agency
Air Quality CO	US Environmental Protection Agency
Air Quality SO_2	US Environmental Protection Agency
Air Quality Ozone	US Environmental Protection Agency
Air Quality Particulate Matter (1.0)	US Environmental Protection Agency
Air Quality Particulate Matter (2.5)	US Environmental Protection Agency
Average Ambient Outdoor Temperature (C)	US Environmental Protection Agency
Outdoor Temperature (F)	US Environmental Protection Agency
Relative Humidity in Air	US Environmental Protection Agency
Dissolved Pb in Water	US Geological Survey
Recoverable Pb in Water	US Geological Survey
Total Pb in Water	US Geological Survey
Mean Soil Pb Measure	US Geological Survey
Median Soil Pb Measure	US Geological Survey
No. Of Scientific Articles per 1000 Graduate Students	National Science Foundation
No. People with Bachelor's Degree in Workforce	National Science Foundation
Federal R&D Expense	National Science Foundation
State R&D Revenue	National Science Foundation
Graduate Students per 1000 of Population	National Science Foundation
State R&D Expense per GDP	National Science Foundation
Venture Capital R&D Expense per GDP	National Science Foundation
Area of Land	US Census Bureau
Number of PFOA Manufacturers	US Environmental Protection Agency
Centroid Latitude	US Census Bureau
Centroid Longitude	US Census Bureau
Republican Campaign Contributions	PEW Research Center
Non-Affiliated Political Party Campaign Contributions	PEW Research Center
Democratic Campaign Contributions	PEW Research Center
National Priority List (Superfund) Sites	US Environmental Protection Agency
EJ Screen Index	US Environmental Protection Agency
No. Of Structures Built After 1970	US Census Bureau - Derived
No. Of Structures Built Before 1970	US Census Bureau - Derived
No. Of Structures Built Before 1960	US Census Bureau - Derived
No. Of Structures Built Before 1950	US Census Bureau - Derived
Population of High School Education or Less	US Census Bureau - Derived
Population of College Education or Higher	US Census Bureau - Derived

Table S-2
Analysis of Variance (ANOVA) Table of final specified models.



```
Analysis of Variance Table

Model 1: y ~ structure_built_before_1950 + sq1950
Model 2: y ~ environment + structure_built_before_1950
Model 3: y ~ indoor + structure_built_before_1950
Model 4: y ~ outdoor + structure_built_before_1950
Model 5: y ~ both + structure_built_before_1950
Model 6: y ~ years + structure_built_before_1950
Model 7: y ~ years + structure_built_before_1950
Model 8: y ~ years + structure_built_before_1950
Model 9: y ~ Y91Y95 + environment + structure_built_before_1950
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     48 1.13376
2     48 0.89929  0  0.234469
3     48 0.98265  0 -0.083360
4     48 0.86389  0  0.118763
5     48 0.92261  0 -0.058720
6     48 0.99133  0 -0.068723
7     48 0.92593  0  0.065400
8     48 0.95872  0 -0.032792
9     47 0.85707  1  0.101656 5.5746 0.02242 *
```

Table S-3
Quadratic effect model residual analysis.

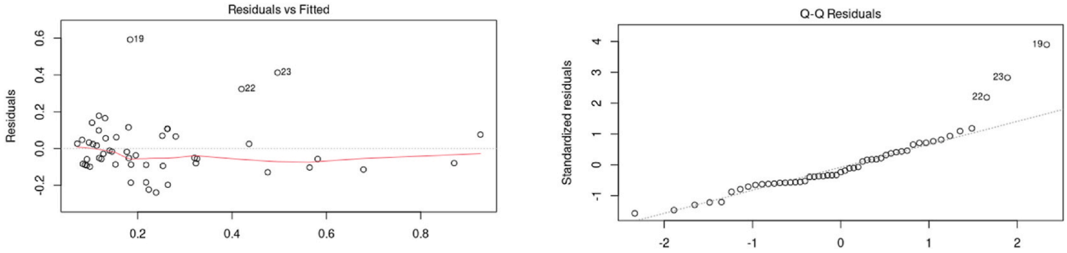
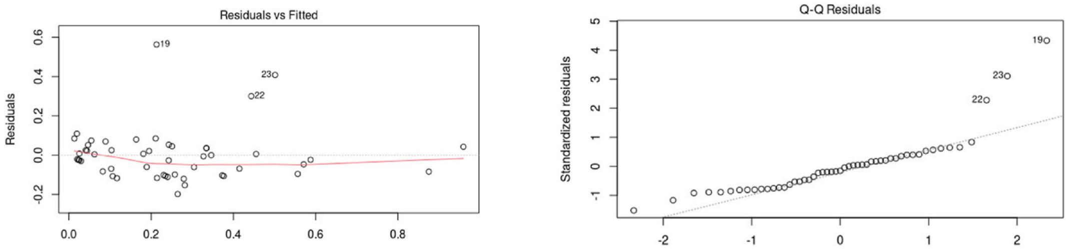


Table S-4
First order model with 1990-95 time-interval and environmental subtopic indicator variable residual analysis.



References

[1] Daniel Kahneman, Thinking, Fast and Slow, 2011. Print.

[2] M. Price, J. Jones, Fitness-maximizers employ pessimistic probability weighting for decisions under risk, Evolutionary Human Sciences 2 (2020) E28, https://doi.org/10.1017/ehs.2020.28.

[3] A. Kappes, A.H. Harvey, T. Lohrenz, R.M. P, T. Sharot, Confirmation bias in the utilization of others' opinion strength, Nat. Neurosci. 23 (1) (2020) 130–137, https://doi.org/10.1038/s41593-019-0549-2.

[4] Kaufman, Kenn "Ruby-throated Hummingbird", National Audubon Society. https://www.audubon.org/field-guide/bird/ruby-throated-hummingbird.

[5] Joseph Millard, Robin Freeman, Tim Newbold, Text-analysis reveals taxonomic and geographic disparities in animal pollination literature, Ecography 43 (2019), https://doi.org/10.1111/ecog.04532.

[6] R. Marks, R. Pauline, Native pollinators, in: Fish and Wildlife Habitat Management Leaflet, 2005. May; Number 34, https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1048334.pdf. (Accessed 30 August 2022).

[7] About Us." United Nations, https://www.un.org/en/about-us#:~:text=The%20United%20Nations%20is%20an,contained%20in%20its%20founding%20Charter. [accessed 30 August 2022].

[8] Jeffrey W. Knopf, Doing a literature review, PS Political Sci. Polit. 39 (1) (2006) 127–132. JSTOR, http://www.jstor.org/stable/20451692. (Accessed 5 October 2022).

[9] P.G. Altbach, H. de Wit, Too much academic research is being published, International Higher Education (96) (2018) 2–3, https://doi.org/10.6017/ihe.2019.96.10767.

[10] Iain J. Marshall, Byron Wallace, Toward systematic review automation: a practical guide to using machine learning tools in research synthesis, BioMed Central 8 (163) (2019).

[11] Cassandra Willyard, Literature reviews made easy, in: American Psychological Association, 2012. https://www.apa.org/gradpsych/2012/03/literature.

[12] Peter H. Westfall, Andrea L. Arias, Understanding Regression Analysis: A Conditional Distribution Approach, CRC Press, 2020.

[13] Frost Jim. Hypothesis testing: an intuitive guide for making data driven decisions, Statistics by, Jim Publishing, 2020, pp. 18–23.

[14] Gui-Quan Sun, Marko Jusup, Zhen Jin, Yi Wang, Zhen Wang, Pattern transitions in spatial epidemics: mechanisms and emergent properties, Phys. Life Rev. 19 (2016) 43–73, https://doi.org/10.1016/j.plrev.2016.08.002. ISSN 1571-0645.

[15] Z. Griliches, The search for R&D spillovers, Scand. J. Econ. 94 (1992) S29–S47, https://doi.org/10.2307/3440244.

[16] HUD (U.S. Department of Housing and Urban Development), American Healthy Homes Survey, American Healthy Homes Survey Lead and Arsenic Findings (Lead Concentration Data provided in 2016 from Policy and Standards Division, Office of Lead Hazard Control and Healthy Homes, U.S. Department of Housing and Urban Development, 2011.

[17] S.O. Teye, J.D. Yanosky, Y. Cuffee, X. Weng, R. Luquis, E. Farace, L. Wang, Exploring persistent racial/ethnic disparities in lead exposure among American children aged 1-5 years: results from NHANES 1999-2016, Int. Arch. Occup. Environ. Health 94 (4) (2021) 723–730, https://doi.org/10.1007/s00420-020-01616-4. Epub 2021 Jan 4. PMID: 33394180.

[18] M. Hauptman, J.K. Niles, J. Gudin, H.W. Kaufman, Individual- and community-level factors associated with detectable and elevated blood lead levels in US children: results from a national clinical laboratory, JAMA Pediatr. 175 (12) (2021 Dec 1) 1252–1260, https://doi.org/10.1001/jamapediatrics.2021.3518. PMID: 34570188; PMCID: PMC8477303.

[19] J. Xue, V. Zartarian, R. Tornero-Velez, L.W. Stanek, A. Poulakos, A. Walts, K. Triantafillou, M. Suero, N. Grokhowsky, A generalizable evaluated approach, applying advanced geospatial statistical methods, to identify high lead exposure locations at census tract scale: Michigan case study, Environ. Health Perspect. 130 (7) (2022) 77004, https://doi.org/10.1289/EHP9705. Epub 2022 Jul 27. PMID: 35894594; PMCID: PMC9327739.

[20] O'Sullivan David, David J. Unwin, Geographic Information Analysis, John Wiley & Sons, Inc., 2010.

[21] "Package 'leaps'", leaps: regression subset selection. https://cran.r-project.org/web/packages/leaps/leaps.pdf. (Accessed 24 January 2023).

[22] "What is Superfund?", in: U.S. Environmental Protection Agency, 2022. November 1, https://www.epa.gov/superfund/what-superfund. (Accessed 21 December 2022).

[23] L.R. Blanco, J. Gu, J.E. Prieger, The impact of research and development on economic growth and productivity in the U.S. States, South. Econ. J. 82 (2016) 914–934, https://doi.org/10.1002/soej.12107.

[24] A.M. Simundić, Bias in research, Biochem. Med. 23 (1) (2013) 12–15, https://doi.org/10.11613/bm.2013.003. PMID: 23457761; PMCID: PMC3900086.

[25] T. Brennan, P. Ernst, J. Katz, E. Roth, Building an R&D strategy for modern times, McKinsey Global Publishing (2020 Nov).

[26] E. Wargo, The mechanics of choice, Association for Psychological Science (2011). Dec 28. Cover Story, https://www.psychologicalscience.org/observer/the-mechanics-of-choice. (Accessed 30 August 2022).