bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A multivariate analysis of CalEnviroScreen: comparing environmental and socioeconomic stressors versus chronic disease

Ben K. Greenfield [1 *], Jayant Rajan [2] and Thomas E. McKone [1 #]

1. Environmental Health Sciences Division, School of Public Health, University of California, Berkeley, 50 University Hall #7360, Berkeley, CA 94720, USA

2. Department of Medicine, Zuckerberg San Francisco General Hospital, University of California, San Francisco, 1001 Potrero Avenue, San Francisco, CA 94110, USA, email: Jayant.Rajan@ucsf.edu

* Corresponding Author. Current Address: Environmental Sciences Department, Southern Illinois University – Edwardsville, IL 62026, USA, email: begreen@siue.edu

email: temckone@lbl.gov

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Background: The health-risk assessment paradigm is shifting from single stressor evaluation towards cumulative assessments of multiple stressors. Recent efforts to develop broad-scale public health hazard datasets provide an opportunity to develop and evaluate multiple exposure hazards in combination.

Methods: We performed a multivariate study of the spatial relationship between 12 indicators of environmental hazard, 5 indicators of socioeconomic hardship, and 3 health outcomes. Indicators were obtained from CalEnviroScreen (version 3.0), a publicly available environmental justice screening tool developed by the State of California Environmental Protection Agency. The indicators were compared to the total rate of hospitalization for 14 ICD-9 disease categories (a measure of disease burden) at the zip code tabulation area population level. We performed principal component analysis to visualize and reduce the CalEnviroScreen data and spatial autoregression to evaluate associations with disease burden.

Results: CalEnviroScreen was strongly associated with the first principal component (PC) from a principal component analysis (PCA) of all 20 variables (Spearman $\rho = 0.95$). In a PCA of the 12 environmental variables, two PC axes explained 43% of variance, with the first axis indicating industrial activity and air pollution, and the second associated with ground-level ozone, drinking water contamination and $PM_{2.5}$. Mass of pesticides used in agriculture was poorly or negatively correlated with all other environmental indicators, and with the CalEnviroScreen calculation method, suggesting a limited ability of the method to capture agricultural exposures. In a PCA of the 5 socioeconomic variables, the first PC explained 66% of variance, representing overall socioeconomic hardship. In simultaneous autoregressive

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

models, the first environmental and socioeconomic PCs were both significantly associated with the disease burden measure, but more model variation was explained by the socioeconomic PCs.

Conclusions: This study supports the use of CalEnviroScreen for its intended purpose of screening California regions for areas with high environmental exposure and population vulnerability. Study results further suggest a hypothesis that, compared to environmental pollutant exposure, socioeconomic status has greater impact on overall burden of disease.

Keywords: health hazard; spatial regression; environmental exposure; socioeconomic status; vulnerable populations; multivariate analysis; principal component analysis; census; environmental justice; CalEnviroScreen

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background

A wide range of factors including demography, socioeconomic status, psychosocial stressors, and environmental exposures influence health outcomes [1–6]. One often used approach for addressing a single stressor is a risk assessment, which focuses on the probability of harm and is based on a quantitative convolution of exposure assessment with a dose-response assessment to provide an overall characterization of risk. But in order to protect vulnerable individuals and communities, environmental health science has broadened in emphasis from single-stressor evaluations to include integrated assessment of multiple stressors [1, 6–10]. This integration among disparate exposures presents a significant methodological challenge, requiring qualitative and less-formal quantitative methods that address hazard as the potential for harm [3, 4, 6, 7, 11, 12]. These impact assessments use metrics of exposure and dose-response but lack the quantitative direct link of these two factors that is common in risk assessment. Based on these methods, environmental justice advocates and health geographers have developed a variety of maps, indices, and tools that integrate environmental health hazards from multiple stressors at varying geographic scales. These tools incorporate a range of indicators including pollutant concentration or load estimates, contaminated sites or other hazards, built environment measures (e.g., urbanization, industry, green space, and road and traffic density), and population characteristics, such as educational attainment and socioeconomic status [2, 6, 8–10, 12–17]. A particular goal is the identification of geographic regions where vulnerable populations encounter environmental exposures, resulting in potential health impacts [2, 16, 18, 19]. To increase transparency and potential societal benefits, many integrated hazard assessment programs also engage the community at large in tool development and assessment, through community-based participatory research, solicitation of public commentary, and the

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

provision and use of publicly accessible data [12, 13, 16, 20–22]. Some of these integration tools and methods also consider the potential interaction between environmental contributors to risk and the preexisting vulnerability of exposed populations to environmental stressors [1, 8, 16, 19, 23].

To complement the development of new methods, there is an ongoing need to quantitatively examine existing methods and tools. Critical and impartial evaluations will ensure that integrated assessment methods are well characterized, technically defensible, and appropriate for intended uses. Further, methods and their underlying data can be analyzed for geographic and statistical patterns of public health hazards [9, 10, 16, 23, 24]. The correlation among and between environmental exposures, socioeconomic vulnerability, and health outcomes also warrant investigation. In particular, an understanding of which health stressors (e.g., environmental, social, economic) are most associated with adverse health outcomes can aid in resource allocation and health policy direction across regions and populations [4]. For example, the relative health impact of environmental factors (e.g., pollution) versus population attributes (e.g., socioeconomic status, educational attainment) warrants examination. Multivariate methods (e.g., ordination, principal component analysis) and spatial statistics [25–27] can be very useful methods for these questions given the multivariate and spatial nature of health hazard assessment [9, 10].

An important case study for spatial health hazard evaluation is the California Communities Environmental Health Screening Tool version 3.0 (hereafter abbreviated as CalEnviroScreen). CalEnviroScreen was developed by the State of California Environmental Protection Agency's (CalEPA) Office of Environmental Health Hazard Assessment (OEHHA) as a "screening

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

methodology that can be used to help identify California communities that are disproportionately burdened by multiple sources of pollution [28].” CalEnviroScreen generates a numeric score, potentially ranging from 0 to 100. The score is based on 20 indicators: 12 measures of environmental exposure, 5 of socioeconomic vulnerability, and 3 of health outcomes (asthma, low birth weight, and cardiovascular disease) [13, 23, 24]. In addition to describing the spatial patterns of hazard, CalEnviroScreen is also intended to help guide state resource allocation. In particular, California Assembly Bill 32 (AB32) and Senate Bill 535 have established a cap and trade program for greenhouse gas emissions, and AB1550, passed in September 2016, requires that 25% of the anticipated one billion dollars of annual state revenue from this program be allocated to communities identified by CalEPA as having health vulnerabilities. CalEPA used CalEnviroScreen to identify these vulnerable communities [18, 23, 28].

The methodology for developing CalEnviroScreen has been detailed elsewhere, and CalEnviroScreen has previously been shown by the scientific development team to indicate strong racial disparities in environmental and socioeconomic vulnerability [23, 24]. The method and some of the underlying assumptions have also been subject to scrutiny as part of a public review process [21, 29], and a previous iteration (CalEnviroScreen version 2.0) was employed as a community disadvantage indicator, and found to be significantly associated with ovarian cancer survival [30]. Despite these applications and their policy significance and an October 10, 2017 search of all peer reviewed journal publications in Google Scholar and in PubMed that include the term “CalEnviroScreen” revealed no independent evaluation of CalEnviroScreen in the peer reviewed scientific literature. This is surprising, given that the

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CalEnviroScreen score will be used to decide on the allocation of around one billion dollars of state revenue, annually.

We examined the multivariate associations in the data underlying CalEnviroScreen (Version 3.0) and the statistical association between the 17 environmental and socioeconomic variables that describe environmental hazard and population vulnerability with an indicator of disease burden. We developed the disease burden indicator using publicly available hospital discharge data at the zip code scale.

Motivation and Specific Aims

The literature on model performance evaluation makes clear the need for systematic efforts to show how a decision-support model performs relative to its intended objectives [31]. The work here was motivated by our observation that in the documentation to date for CalEnviroScreen, there is limited information regarding the multivariate associations in the variables used to construct the overall impact score. For example, there is no comparison to separate burden of disease measures, and no examination of the relationship between these associations and the CalEnviroScreen scores, themselves. Analyzing thirty sites at a zip code scale using a preliminary variable set and model formulation in 2012, Meehan August et al. [13] found correlations among many indicator variables proposed for inclusion in CalEnviroScreen, and low to moderate score sensitivity to changes in CalEnviroScreen model formulation. However, since that time the model formulation, variables, data sets, and geographic resolution have all changed. The current CalEnviroScreen model provides census tract scale data for over 8000 sites.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The CalEnviroScreen data reduction and score calculation methodology is moderately complex, and is based on considerable professional judgment [13, 23]. Public review comments frequently raise concerns that this calculation method may distort the comparative differences in hazard among locations, suggesting that comparison to an independently developed aggregate score would be useful. Therefore, one of our study objectives was to perform an external validation of the CalEnviroScreen score by comparison to a score obtained using simple application of a multivariate statistical algorithm (principal component analysis) to the underlying variables. We confront the need to quantitatively address outcome dependence on inputs in order to provide insight for decision makers.

In short, our aim was to systematically to describe associations among variables that go into the CalEnviroScreen score, whether the score strongly correlates with systematic change in the combined variables, and whether these indicators of hazard and vulnerability predict burden of disease. To achieve these objectives, we addressed four questions about the CalEnviroScreen source data and its use to produce an outcome score: 1. What are the correlation patterns and statistical associations in the variables? 2. How does the CalEnviroScreen numeric score compare to a score based on principal component analysis, employed as an alternate scoring method based on the statistical associations in the underlying data? 3. Using the alternate scoring method, what spatial patterns are evident in environmental hazard and population vulnerability in California? 4. Do the environmental hazard and population vulnerability indicators statistically predict overall disease burden? For this last question, we work to generate hypotheses regarding the ability of CalEnviroScreen to address health outcomes and the relative importance of environmental versus socioeconomic factors in determining disease burden.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

CalEnviroScreen background and data

We downloaded the CalEnviroScreen data along with the CalEnviroScreen 3.0 scores as a Microsoft Excel spreadsheet file from the CalEnviroScreen website [32]. These data have been pre-cleaned and carefully prepared by CalEPA, as described elsewhere [23, 24]. The data set covers 8035 census tracts, though CalEnviroScreen 3.0 results are calculated for 7929 census tracts. CalEnviroScreen 3.0 includes 12 environmental hazard variables: ozone levels, concentrations of particulate matter $\leq$2.5 μm in diameter (hereafter, $PM_{2.5}$), diesel particulate matter concentrations (diesel PM), traffic density, drinking water contamination, active pesticide mass used in agriculture (pesticide use), airborne toxic chemical releases, water body impairments, sites hazardous to groundwater (groundwater threats), sites targeted for cleanup, hazardous waste sites, and solid waste sites. CalEnviroScreen 3.0 also includes 5 socioeconomic vulnerability variables: low educational attainment, linguistic isolation, poverty, unemployment, and households severely burdened by housing costs (hereafter, housing burden). CalEnviroScreen 3.0 additionally includes three specific health outcome variables: asthma, low birth-weight, and cardiovascular disease, which are intended to indicate a combination of vulnerability to, and effects of, environmental exposures [23]. All variables were obtained based on data collected between 2009 and 2016, except drinking water contamination (2005 – 2013). Table S1 (Additional File 1) summarizes the variables, providing abbreviations, years represented, original units, and data transformations for this study. Cushing et al. [24], Faust et al. [23], and OEHHA [33] provide more extensive detail.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

178 The CalEnviroScreen 3.0 score is constructed from these data by a weighted averaging and
179 multiplication algorithm, described elsewhere [23]. Briefly, each underlying variable is
180 assigned a percentile rank score across each available census tract, a single weighted average is
181 calculated for the 12 environmental variables and another weighted average for the 8
182 population variables (socioeconomic vulnerability and health outcome variables), each
183 weighted average is linearly rescaled from 0 to 10, and finally, the two averages are multiplied
184 achieving a score possibly ranging from 0 to 100.

185 **Disease burden measure**

186 Although CalEnviroScreen does incorporate three specific health outcomes, we developed a
187 separate disease burden measure independently of CalEnviroScreen to examine how well the
188 environmental and socioeconomic variables in CalEnviroScreen predict overall burden of
189 disease. To maintain transparency and potential for community access [4, 13], our disease
190 burden indicator was developed using publicly available hospital discharge data collected at the
191 zip-code level. Although some researchers may not consider hospitalization at zip code as the
192 most specific measure that one would like for a disease burden analysis, it is what is publicly
193 available to derive metrics of disease burden and has been used in other studies tracking links
194 between environmental quality and disease [34, 35].

195 We characterized burden of diseases using discharge diagnostic codes (which used the ICD-9-
196 CM schema). We obtained these data for all hospitalizations for a given calendar year using
197 publicly available, de-identified, statewide hospital discharge data from the California Office of
198 Statewide Health Planning and Development, spanning the years 2008-2011. Using these data,
199 we classified hospitalizations by pre-determined ICD-9 diagnostic categories, focusing on

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

diseases having environmental etiology. We determined the sum total number of hospitalizations for 14 disease diagnostic categories representing serious or chronic ailments known to have potential environmental etiology. The 14 categories were pneumonia, chronic obstructive pulmonary disease (COPD), asthma, myocardial infarction (MI), cerebrovascular accident (CVA), diarrhea, pancreatic cancer, lung cancer, breast cancer, lymphoma, leukemia, depression, schizophrenia, and low birth weight.

The overall approach of combining multiple diseases without weighting factors and assuming that hospital admissions can be a metric of regional disease patterns offers utility for impact assessment but has limitations. Combining multiple disease outcomes into a single metric of disease burden has the advantage of being comprehensive in capturing a large range of environmentally related diseases. The approach has been used in assessing global burden of disease for policy studies [36], but how or whether to use disease-weighting factors is a concern [37], and selection of weighting factors can both enhance and bias results. Because we were not interested in a metric of disability but only wide-scale disease patterns, we elected not to use disease-weighting factors. Using hospital admissions data as a metric of regional variations in disease patterns also has utility a health metric but limitations with regard to how well this metric relates to the health of a community. These issues have been addressed in the health tracking literature and show that the approach can be useful to policy makers [34, 38].

In a preliminary analysis, we found that the occurrence of these disease categories in hospitalized persons was positively correlated among the categories and also correlated with the total number of diagnoses (Additional File: Table S2, Fig. S1). Further, because the same patient hospitalization could have multiple diagnoses spanning multiple categories, rates for

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

different disease categories are not mutually exclusive. To address these issues, and minimize the Type I error rate and complexity of the analysis, we assembled these data into a single count of total hospitalizations resulting from the 14 diagnostic categories. In this count, a hospitalization event that included more than one diagnostic category was recorded as a single event.

We summed the count of total hospitalizations at the zip code tabulation area (ZCTA) level, and divided by total population. For the denominator, we used ZCTA population estimates from the 2010 United States Census. We observed high variability in rates for populations below 100 individuals within a ZCTA (Additional File: Fig. S2). Therefore, we excluded ZCTAs having populations < 100 individuals from the final analysis to minimize the influence of statistical outliers.

In summary, this disease burden measure is the total rate of hospitalizations associated with at least one of the 14 diagnostic categories selected for a given ZCTA. Because the same person could in theory be admitted multiple times for the same diagnosis, the rates reported here are only approximations of population disease prevalence. They may be viewed as representing the impact of certain diseases (particularly chronic diseases), since apart from mortality, hospitalization is generally the most extreme result of any disease process.

Expecting chronic disease burden to be higher among the elderly, we obtained the percent of the population over 65 years old for each ZCTA from the 2010 US Census. This parameter (Over65) was used in addition to the environmental and socioeconomic variables derived from CalEnviroScreen as a predictor in statistical models used to explain the disease burden measure.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data preparation and spatial alignment

Hospitalization-related diagnoses were tabulated at the zip code scale, census population is at the ZCTA scale, and CalEnviroScreen data are available at the census-tract scale. Thus, spatial alignment was required prior to comparing CalEnviroScreen variables to the disease burden measure. The following data preprocessing steps were performed using ArcMap v10 (ESRI, Redlands, CA): 1. average the CalEnviroScreen data at the zip code scale to achieve a consistent analysis scale; 2. link the CalEnviroScreen and ICD-9 data together; 3. combine with a shapefile of zip code polygons; and 4. standardize the health outcome data to total population per zip code tabulation area (ZCTA). Additional File 1 further details this spatial alignment methodology.

Statistical methods

Statistical analysis was performed in R (version 3.4.0) [39]. Principal component analysis (PCA) was conducted for data reduction, and to allow examination of the contribution to variance explained by individual variables (i.e., variable loadings) and the multivariate data structure, similar to other public health studies [10, 40, 41]. For all census tracts that had CalEnviroScreen results (n = 7929), PCA was performed on the correlation matrix using the R package factoMineR. Prior to PCA, missing values were imputed using the imputePCA command in the missMDA package [42].

Separate PCAs were performed to achieve different study objectives. A PCA was performed on all 20 CalEnviroScreen variables in combination in order to examine multivariate patterns of the entire data set. Two separate PCAs were also performed on the 12 environmental and the 5

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

265 socioeconomic variables to generate and evaluate a smaller number of variables representing
266 the categories of environmental hazard, and socioeconomic status, respectively. Finally,
267 another PCA was performed on the 17 environmental and socioeconomic variables. This PCA,
268 which did not include the three health outcome variables (asthma, low birth-weight, or
269 cardiovascular disease), was compared to the hospitalization rate disease burden measure
270 (described above). The goal of this analysis was to examine how these exposure and population
271 variables underlying CalEnviroScreen are generally associated with disease burden.

272 Environmental hazard and socioeconomic status variables (principal components) were
273 compared to the disease burden measure (hospitalization rate for 14 diagnoses) using
274 simultaneously autoregressive models (SAR), employing the R package spdep. SAR was
275 chosen as a spatial autoregressive model appropriate to describe and test for linear relationships
276 in the presence of spatial autocorrelation [27, 43]. Appropriate treatment of spatial
277 autocorrelation was assessed based on Moran plots illustrating no association with spatially
278 lagged means, global Moran's I that was not significant, and a spatial dependence parameter (λ)
279 that was significant via likelihood ratio test [43–45]. Models were selected based on
280 minimizing the Bayesian Information Criterion (BIC). Parameter inclusion was based on
281 reported p values ($\alpha = 0.05$) and on ΔBIC, employing the rule of thumb that $\Delta BIC \geq 2$ provides
282 positive evidence of model improvement [46]. Nagelkerke pseudo-R^2 was calculated as a
283 measure of model goodness of fit for SAR models. Analogous to traditional R^2 in meaning
284 (though not directly comparable), the Nagelkerke pseudo-R^2 estimates from 0 to 1 the
285 improvement in proportion of variation explained by the fitted model, versus a null (intercept-
286 only) model [47]. In order to compare the contribution of each parameter to final variation

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

287 explained by the model, the psuedo-R^2 was compared between the full model and the model
288 with that parameter removed.

289 Prior to statistical analysis, all variables were transformed to approximate a normal distribution
290 and multivariate linearity required for linear model analysis [26, 48]. Transformations included
291 $\log_{10}$ (7 variables), cube root (6 variables), square root (5 variables), and arcsine square root
292 transformation (drinking water). $PM_{2.5}$ and low birth weight did not require transformation
293 (Additional File: Table S1). The combined disease burden measure (DB) exhibited skewness
294 and long tails (leptokurtic) and standard transformations failed to achieve normally distributed
295 model residuals. Normal residuals were therefore achieved employing a modulus
296 transformation: $\text{sign}(\sqrt{\text{DB}}) * \ln(|\sqrt{\text{DB}}| + 1)$ following John and Draper [49]. The predictor
297 variables for the SAR were centered and scaled by subtracting the mean and dividing by the
298 standard deviation. This converted the transformed variables (Additional File: Table S1) to the
299 same unit normal distributions, such that a comparison of model coefficients would
300 approximately indicate relative contribution of each variable to disease burden [50].

301 In the interest of independent assessment, we did not communicate with OEHHA, CalEPA, or
302 any members of the CalEnviroScreen development team regarding any aspect of this study.

303 ## Results

304 For Question 1, describing the multivariate structure of the CalEnviroScreen source data, we
305 employ PCA to visualize which exposures are associated with each other and the prevailing
306 patterns of overall exposure encountered in California. For Question 2, how CalEnviroScreen
307 compares to an alternate metric based on PCA, we present the correlation between the main

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

principal components and the CalEnviroScreen score. For Question 3, spatial patterns in environmental hazard and population vulnerability, we map the first principal component (PC) of separate PCAs performed on the environmental hazard and socioeconomic vulnerability variables. Finally, Question 4 examines and compares whether the exposure and vulnerability indicators in CalEnviroScreen predict disease burden. We employ the PCs rather than the individual parameters to focus on overall patterns of exposure and vulnerability and to reduce the number of required analyses. The relative importance of environmental versus socioeconomic parameters in the model illustrates which factors most influence disease burden.

Multivariate data structure

Pearson's pairwise correlation coefficients (r) indicate multiple associations for the underlying CalEnviroScreen data (Table 1). Positive pairwise associations are observed among variables related to air pollution and traffic, with diesel PM moderately correlated with $PM_{2.5}$ (r = 0.41) and traffic density (r = 0.56), and toxic release correlated with these three variables (r = 0.43 – 0.56, Table 1). Socioeconomic indicators of vulnerability are also positively associated: low educational attainment, linguistic isolation, poverty, and unemployment exhibit r values ranging from 0.51 to 0.82, with the exception of linguistic isolation versus unemployment (r = 0.24) (Table 1; Additional File: Fig. S3). Housing burden is also positively related to these variables, with correlation coefficients ranging from r = 0.40 versus unemployment to r = 0.72 for poverty. The strongest correlation between socioeconomic and environmental variables is between linguistic isolation and diesel PM (r = 0.43). The strongest negative association among all variables is groundwater threats versus ozone (r = −0.33). Of the CalEnviroScreen health outcome variables, low birth weight is not well correlated with any environmental variables,

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

but is highly correlated with housing burden (r = 0.72). Similarly, asthma is more positively correlated with the socioeconomic variables education (r = 0.51), poverty (0.53), and housing burden (0.50), than with any environmental variables. Cardiovascular disease exhibits moderate correlations with low educational attainment, poverty, and unemployment (r = 0.43 – 0.46) but also with the environmental variable, ozone (0.39).

<Insert Table 1 Here>

Principal component analysis and comparison to CalEnviroScreen

PCA was performed on the entire CalEnviroScreen data set (20 variables), on the data set without the health outcome variables (17 variables), and on the environmental (12 variables) and socioeconomic (5 variables) data. For the entire data set, the first three PCs explain 50% of data variability in combination. The first principal component (PC) (Fig. 1a, horizontal axis), explaining 23% of variation, is positively associated with all variables except for a weak negative association with pesticide use and impaired water bodies. The variables with the greatest variance along this axis are asthma (health indicator) and the five socioeconomic indicators (linguistic isolation, low educational attainment, poverty, unemployment, and housing burden). The other two health indicators (low birth weight and cardiovascular disease) are also associated with the first PC axis, as is $PM_{2.5}$ and diesel PM.

The second PC (Fig. 1a, vertical axis; Fig. 1b, horizontal axis) explains 14% of variation. Indicators of industrial pollution and associated hazardous sites score positively with both the first PC (hereafter referred to as PC1all) and the second PC (hereafter, PC2all) (Fig. 1a, b). In particular, PC1all and PC2all are associated with groundwater threat sites, hazardous waste

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sites, cleanup sites, toxic release, traffic, and diesel PM. These variables are negatively correlated with pesticide use, which would be expected in rural areas, as well as drinking water contamination and ozone. Examining the biplot of PC2all and PC3all (Fig. 1b), we see a negative association between the polluted sites (cleanup sites, groundwater threats, impaired water bodies, and hazardous waste sites) and measures of drinking water contamination and ozone. Additionally, motor vehicle and industrial source-associated air pollution indicators (traffic, diesel PM, and toxic release) are negatively associated with pesticide use. Results for the PCA of all except the health outcome variables (17 remaining variables) are qualitatively very similar to that of the entire data set.

For the environmental PCA (Fig. 1c), the first two principal components explain 43% of data variability in combination. Along the axis of the first PC, explaining 23% of variance, there is a positive association among toxic releases and motor vehicle pollution indicators: $PM_{2.5}$, diesel PM, and traffic. These are negatively associated with pesticide use. Sites contaminated due to industrial activity, including groundwater threats, hazardous waste sites, and cleanup sites, are positively associated with both the first and second PCs. Based on these associations, the first PC (hereafter, PC1env) represents general exposure to urban and industrial pollution. Interestingly, PC1env is strongly positively correlated at the zip code scale with population density ($r = 0.81$, $n = 1602$), indicating densely populated areas are more exposed to the main environmental hazards measured by CalEnviroScreen.

The second environmental PC (hereafter, PC2env) explains 20% of variance, and is negatively associated with ozone and drinking water contamination (Fig. 1c). Examination of the associations of PC2env indicates that elevated hazards due to ozone and drinking water

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

373 contamination will tend to occur in different areas from impaired water bodies or groundwater
374 threats.

375 For the socioeconomic PCA (Fig. 1d), the first two PCs explain 82% of data variance. The first
376 PC (hereafter PC1soc) explains 66% of variance and is positively associated with all five
377 socioeconomic indicators. Thus, PC1soc broadly indicates socioeconomic vulnerability. The
378 second PC (PC2soc) explains 16% of variance, and separates unemployment from linguistic
379 isolation (Fig. 1d).

380 The CalEnviroScreen method produces a numeric score calculated as a weighted sum of
381 environmental variables (which CalEPA refers to as pollution burden) multiplied by a weighted
382 sum of the socioeconomic and health outcome variables (referred to as population
383 characteristics) [23, 24]. We were interested in how well this calculation method represents the
384 prevailing multivariate patterns in environmental and socioeconomic vulnerability within the
385 population of census tracts. That is, does CalEnviroScreen achieve its intended purpose of
386 identifying areas exhibiting high hazard from a combination of environmental exposures and
387 socioeconomic vulnerabilities? To evaluate the validity of CalEnviroScreen based on this
388 criterion, we calculated Spearman's rank correlation coefficient for the CalEnviroScreen score
389 versus the main PCs from each PCA analysis, for all available census tracts (n = 7929).
390 Spearman's ρ was employed as a measure of association that is robust to nonlinear
391 relationships, which were evident in this analysis (Fig. 2).

392 In the all-variables PCA, CalEnviroScreen is strongly associated with PC1all ($\rho = 0.95$; Fig. 2)
393 and not associated with PC2all ($\rho = 0.00$) or PC3all ($\rho = 0.10$). When separate PCAs are
394 performed on the environmental and socioeconomic data, CalEnviroScreen is strongly

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

associated with PC1soc ($\rho = 0.82$), moderately associated with PC1env ($\rho = 0.59$), and not associated with PC2env ($\rho = -0.08$) or PC2soc ($\rho = 0.00$). These results indicate that changes in the predominant gradients underlying the data (PC1all, PC1soc, and PC1env) are generally captured by the CalEnviroScreen score. Thus, this single score effectively captures the prevailing gradients in the underlying variability in environmental exposures and socioeconomic vulnerabilities.

The only variables negatively associated with PC1all (Fig. 1a) were pesticides and impaired water bodies. Not surprisingly, the CalEnviroScreen 3.0 score exhibits no correlations with these two variables; Pearson's r was 0.05 for pesticides and −0.04 for impaired water bodies. As a result, the CalEnviroScreen census tract rankings will be insensitive to these two variables.

Spatial patterns

The spatial patterns in environmental hazard and population vulnerability can be seen by plotting the first principal component for the environmental and socioeconomic variables, respectively (Fig. 3, Fig. 4). For ease of visualization, we present these results aggregated at the zip code scale; spatial patterns in individual variables and calculated linear averages based on the CalEnviroScreen method are displayed at the census tract scale by the original authors [23]. Strikingly, both PC1Env and PC1Soc exhibit relatively high values in the southern Central Valley region of the state, especially the southwest San Joaquin Valley, indicating a pattern of elevated hazard and vulnerability in this region. PC1Env is also high in the more densely populated San Francisco Bay Area and Los Angeles regions, illustrating the abovementioned association between population density and the environmental hazard variables. In contrast to

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PC1Env, PC1Soc exhibits considerable heterogeneity in both the San Francisco and Los Angeles regions. All of these patterns were similarly observed using the CalEnviroScreen methodology for deriving their pollution burden and population characteristics metrics [23].

Spatial patterns in the CalEnviroScreen 3.0 measure (Fig. 5) clearly combine these two factors. As above (Fig. 2), the complex algorithm employed for aggregating the 20 variables in CalEnviroScreen [23] essentially captures the main underlying environmental and socioeconomic gradients (Fig. 3, Fig. 4). The highest scoring (and thus most impacted) areas are centered around the southwest San Joaquin Valley, peaking in the urbanized portions of Fresno County (including the cities of Fresno and Selma), as well as the Los Angeles region (Los Angeles, Pomona, and San Bernardino).

Comparison to the disease burden measure

The principal components from the all environmental and socioeconomic variables PCA (PC1all, PC2all, and PC3all) and from the separated environmental and socioeconomic PCAs (PC1env, PC2env, PC1soc, and PC2soc) were evaluated as possible predictors for the disease burden measure that we developed. The predictor variables were not correlated with each other ($|r| \leq 0.13$), with the exception of a weak negative correlation between PC1env and PC1soc ($r = 0.31$) and a weak correlation between PC1env and PC2soc ($r = -0.31$). Percent population > 65 years old (hereafter, Over65) was also included as a potential predictor. Over65 was moderately correlated with PC1all ($r = -0.45$), PC1soc ($r = -0.36$), and PC1env ($r = -0.42$), and uncorrelated with the other parameters.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Initial modeling using linear regression to predict disease burden indicated clear evidence of spatial autocorrelation of residuals for all models (based on Moran's plots and significant Global Moran's I). Thus, SAR was used to predict disease burden, and addressed these issues. For the SAR model on the all environmental and socioeconomic data PCA results (n = 1606 Zip Code Tabulation Areas), the following model form was obtained (coefficient estimate ± standard error, SE):

Disease burden = (0.47 ± 0.031)PC1all + (0.59 ± 0.022)Over65

For both PC1all and Over65, ΔBIC > 2, indicating that addition of each parameter was important in explaining disease burden, and $p < 0.0001$. In contrast, PC2all ($p = 0.10$), PC3all ($p = 0.96$), and the intercept ($p = 0.97$) were not significant as added model terms. For the final model, the Nagelkerke pseudo-R^2 (hereafter, R^2) was 0.49, which was effectively unchanged when attempting to include PC2all or PC3all. Change in R^2 after removing individual parameters illustrates the contribution of each parameter to explaining the variability in disease burden. Decrease was greater when removing Over65 (new $R^2 = 0.25$) than when removing PC1all (new $R^2 = 0.42$). The association with PC1all generally describes the association of disease burden with the correlated variability in all of the CalEnviroScreen environmental and socioeconomic indicators except for pesticide use and impaired water bodies (Fig. 1a).

For the separate environmental and socioeconomic PCA results (N = 1606), three model terms contributed to describing disease burden (coefficient estimate ± SE):

Disease Burden = (0.35 ± 0.035)PC1env + (0.31 ± 0.024)PC1soc + (0.21 ± 0.026)PC2soc + (0.60 ± 0.021)Over65

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

458 ΔBIC > 2 and individual parameter p-values were < 0.001 for all included model terms. Neither
459 PC2env (p = 0.16) nor the intercept (p = 0.93) were significant. The model coefficient was
460 largest for Over65, simply indicating that age must be accounted for in the analysis. The
461 positive coefficient for PC1env and PC1soc indicate an overall association with environmental
462 exposures and socioeconomic vulnerability. The coefficient for PC2soc further suggests an
463 association with unemployment, rather than linguistic isolation, in particular. The association
464 with two socioeconomic principal components but just one environmental principal component,
465 with qualitatively similar coefficient magnitudes on standardized variables, suggest a greater
466 importance of socioeconomic than environmental factors in predicting overall disease burden.
467 Similarly, the Nagelkerke pseudo-R^2 indicates more variability explained by socioeconomic
468 variables. Compared to the full model (R^2 = 0.51), the R^2 slightly declined when removing
469 PC1env (R^2 = 0.48), PC1soc (R^2 = 0.46), or PC2soc (R^2 = 0.49) from the model, but declined
470 more when removing Over65 (R^2 = 0.27), or PC1soc and PC2soc in combination (R^2 = 0.44).
471 These results suggest that whereas environmental hazard and socioeconomic vulnerability both
472 contribute, socioeconomic vulnerability is more important than environmental hazard for
473 explaining disease burden at the zip code scale.

474 # Discussion

475 ### Multivariate analysis of the California Communities Environmental Health Screening
476 ### Tool

477 Our study results support the use of CalEnviroScreen as a census tract scale indicator of
478 environmental health hazard and vulnerability. The CalEnviroScreen numeric score was
479 strongly associated with the first principal components in all analyses, indicating that it

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

480 represents the primary underlying gradients within the data set. Additionally, the spatial
481 patterns in the first principal components matched those of the CalEnviroScreen combined
482 indicators and numeric score. Further, the principal components from the CalEnviroScreen
483 exposure and vulnerability variables were significantly associated with our estimate of disease
484 burden, which is a general indicator of health care burden. Models that contained these
485 variables, and also accounted for spatial autocorrelation and proportion of population that was
486 over 65, explained approximately 50% of the variation in the underlying data. Given all of the
487 uncertainty and assumptions with the study data and scale, this suggests that the
488 CalEnviroScreen data includes multiple exposure hazards and socioeconomic vulnerability
489 indicators, which in combination influence burden of disease. Our analysis, therefore, suggests
490 that CalEnviroScreen is an appropriate tool for its intended purpose: to identify vulnerable
491 communities for resource allocation in environmental health restoration [18].

492 The first few principal components explained limited variance in the underlying data set, and
493 many of the parameters, especially environmental exposure measures, were weakly correlated.
494 Similar to our results, Lalloué et al. [9] observed 55% of total variance explained in the first
495 three components of a multiple factor analysis of 17 environmental indicators and Messer et al.
496 [10] observed 23 to 46% of variability explained in the first component in five related domains
497 (air, water, land, built, and sociodemographic). These similar observations of low to moderate
498 variance explained in the first principal components suggest that it is not possible to explain
499 much variability in environmental health hazard using a small subset of indicators. Rather, a
500 complete accounting of environmental hazards and population vulnerability requires a wide
501 range of indicators, as employed in CalEnviroScreen and other studies [9, 10, 12–14, 23].

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The limited variability explained by the first few principal components further suggests that for the 20 hazard parameters captured in CalEnviroScreen 3.0, there will be many exposure, vulnerability, and health outcome combinations that are not fully described by combined multivariate gradients. For example, the negative association of ozone air pollution with groundwater threats and water body impairments is not readily explained but suggests that residents of different regions encounter different exposure hazards. Further examination of the statistical properties and demographic vulnerability of sites exhibiting unique exposure combinations is ongoing. These efforts, performed by local agencies, interest groups, and community-based organizations in evaluating updates to the CalEnviroScreen method [51], complement the CalEnviroScreen numeric score by establishing a typology of vulnerable communities. How to incorporate these efforts into resource allocation decisions remains a difficult policy challenge.

The evaluation and modification of CalEnviroScreen is reflected in the recent release of CalEnviroScreen 3.0, updating much of the data, and attempting to address prior community review comments. In comparison to CalEnviroScreen 2.0, CalEnviroScreen 3.0 added two new indicators (hospital visits for heart attack, and low income households burdened by high housing costs), removed a vulnerable age indicator, and retained the other 18 indicators included in this paper [23]. Despite these changes, the multivariate patterns we observed in the correlation coefficients and principal component analysis results were almost identical between the two versions, and we therefore chose to focus on CalEnviroScreen 3.0 in this paper. Continued examination of the statistical properties and association with health outcomes is warranted for this and other multivariate hazard screening methods [9, 16, 17, 40], to complement the ongoing public discourse and review.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Some policy interventions may best be geographically targeted using additional information beyond the CalEnviroScreen score itself. Most of the environmental hazards measured were associated with urbanization and industrial activities, and the first environmental PC was strongly associated with population density. Although environmental exposure hazards are often elevated in urban areas [10], this is not always the case. For example, agricultural pesticide exposure is linked to a variety of developmental and health effects [52–54]. The pesticide use indicator in the CalEnviroScreen metric largely derived from agricultural application [23], was uncorrelated or negatively associated with the other environmental hazard indicators, and uncorrelated with the metric itself. Elevated nitrate concentrations in drinking water, another agricultural pollutant, is an additional known hazard for Latino communities in the San Joaquin Valley [55]. Given that rural communities tend to have lower incomes and reduced access to medical care [10, 56], additional measures of exposure hazard and vulnerability that affect isolated populations may warrant consideration alongside the CalEnviroScreen score.

Disease burden was more associated with socioeconomic status than environmental hazards

We employed multivariable analytical methods to separate out chemical pollutant exposure hazard versus socioeconomic variation within California. Although environmental hazards and socioeconomic vulnerabilities are often correlated through complex causal pathways, and low income communities often face disproportionate environmental exposures [17], our multivariate approach allowed for a direct comparison of the statistical effect of environmental hazard versus socioeconomic vulnerability indicators on burden of disease. This is because the

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

547 principal components for environmental variation (PC1env, PC2env) and for socioeconomic
548 status (PC1soc, PC2soc) were uncorrelated with each other.

549 The association of the disease burden variable with both socioeconomic principal components
550 but only one environmental principal component, as well as the greater combined contribution
551 to variability explained (change in R^2) suggest a stronger association of disease burden with
552 socioeconomic status than with environmental pollution exposure. This was also evident in the
553 greater correlation of individual socioeconomic variables with cardiovascular disease, asthma,
554 and low birth weight. This supports the paradigm that underlying population vulnerability,
555 resulting from socioeconomic conditions, must be considered in health risk assessment [1, 3,
556 12]. This finding is further acknowledged in that investigations of the environmental causes of
557 disease typically adjust for indicators of socioeconomic status [5, 17]. The generalizability of
558 our finding that socioeconomic factors better explained disease burden than environmental
559 hazards merits investigation, as it would have implications for intervention priorities, as well as
560 for conceptualization of the primary structuring factors that influence disease.

561 The multivariate and exploratory approach of our study reflects objectives quite different from
562 a traditional epidemiological evaluation of how one or a small number of exposures affects a
563 single outcome. Instead, our approach falls within the realm of quantitative methods for
564 comparing among and evaluating cumulative environmental impacts in combination [6, 9, 10,
565 17]. We identified prevailing gradients of exposure and vulnerability, and observed how these
566 patterns were associated with disease burden. We observed relatively strong associations
567 among all of the socioeconomic indicators (education, income, unemployment, linguistic
568 isolation), each of which may exhibit a separate impact on vulnerability [13]. This could

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

569 explain the stronger association between socioeconomic indicators and disease burden, in
570 contrast to environmental hazards, which were less correlated, such that the gradients in
571 multivariate exposures were weaker. In other words, our findings support the paradigm that
572 population disease burden will be more strongly impacted when multiple stressors occur in
573 combination. As such, examination of the multivariate association among stressors should
574 provide added and complementary information to bivariate analyses of exposure versus
575 outcome.

576 **Limitations and caveats**

577 Like all census-tract-scale studies of publicly available spatial exposure and health data, this
578 study has limitations. For their similar study of the San Joaquin Valley region, Huang and
579 London [14] eloquently describe the limitations of studies using publicly-available spatial
580 exposure data. Our study does not establish causality and we cannot extrapolate inferences to
581 the individual level [57]. We used publicly available health outcome data; as such our analysis
582 was restricted to hospital discharge data at the zip code scale, which can be used to indicate
583 overall morbidity [38], but may also be subject to bias [58]. Moreover, the specific choices we
584 made regarding ICD-9 endpoints that had environmental etiology could be questioned, and
585 must be interpreted as a general burden of disease measure, rather than indicative of any
586 specific health outcome. Data required geographic alignment, including assembly of different
587 parameters provided at multiple and varying spatial scales. In particular, CalEnviroScreen data
588 were available at the census-tract level, the disease burden measure at the USPS zip code level,
589 and spatial polygon arrangement at the zip code tabulation area-level. Inaccuracies are
590 inevitably introduced when aligning these different spatial scales [59]. In line with the

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

591 protection of individual rights to anonymity in publicly accessible outcome data, individual-
592 level demographic information was masked, and residential addresses were limited to USPS zip
593 code. These factors likely in part explain the limited strength of associations observed in this
594 study. In particular, the association between the environmental principal components and
595 disease burden was weak in our study. This was similar to Gaffron and Niemeier [19], who
596 observed a very low bivariate strength of association ($R^2 = 0.018$) between $PM_{2.5}$
597 (environmental variable) and emergency visits for asthma (health outcome) employing
598 CalEnviroScreen census tract data in six Sacramento, CA region counties. Especially, studies
599 are needed combining longitudinal data sets of disease occurrence with CalEnviroScreen and
600 other hazard measures [30].

601 ## Conclusions

602 Previous studies at similar spatial scales and resolutions have established relationships of
603 environmental hazards and disease risk with race and socioeconomic status, with implications
604 for resource allocation and policy [14, 15, 19, 24, 60]. Prior studies have also shown
605 geographic indicators of socioeconomic status or vulnerability to be associated with
606 hospitalization rates or disease occurrence [2, 30]. However, few studies explicitly evaluate and
607 describe the multiple patterns of association that occur across a range of health hazards and
608 vulnerabilities at the census-tract scale [9, 10]. We use this approach to evaluate a cumulative
609 impact screening methodology (CalEnviroScreen), observing that the methodology produced a
610 spatial data set that captures the main underlying gradients, which in turn are associated with
611 the burden of diseases having environmental etiology. CalEnviroScreen should therefore be

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

612 useful for its intended purpose of screening for those communities most vulnerable to
613 environmental exposure.

614 We observed that socioeconomic indicators were associated with each other and contributed to
615 explaining disease burden, and that an environmental gradient of urban and industrial pollution
616 also contributed to explaining disease burden. Ground-level ozone and drinking water threats
617 were negatively associated with impaired water bodies and groundwater threats. Some of these
618 findings corroborate findings from the preliminary analysis for CalEnviroScreen development
619 using 30 zip codes by Meehan August et al. [13]. The existence of separate gradients of
620 environmental hazard and socioeconomic disparity, and the varying ability to predict disease
621 burden highlight the need for continued emphasis on integrated approaches in vulnerability
622 assessment.

623 ## Additional files

624 **Additional file 1:** Supporting information (PDF, 11 pages). Includes the following: methods
625 for spatial joining of CalEnviroScreen versus hospital data; correlation of variables in ICD-9
626 hospital data (text and Table S2); study variables description (Table S1); scatterplot matrix of
627 hospital visit rate parameters (Fig. S1); total hospital ICD-9 codes vs. total population (Fig.
628 S2); and scatterplot matrix of socioeconomic variables (Fig. S3).

629 **Abbreviations**

630 Bayesian Information Criterion, BIC; California Communities Environmental Health Screening
631 Tool Version 2.0, CalEnviroScreen; California Environmental Protection Agency, CalEPA;
632 Chronic obstructive pulmonary disease, COPD; Cerebrovascular accident, CVA; Disease

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

633 burden measure, DB; International Classification of Diseases Ninth Revision Clinical
634 Modification, ICD-9-CM; Myocardial infarction, MI; Principal component, PC; Principal
635 component analysis, PCA; Particulate matter, PM; Particulate matter 2.5 μm in diameter or
636 below, $PM_{2.5}$; Pearson's pairwise correlation coefficient, r; Simultaneously autoregressive
637 model, SAR; Zip code tabulation area, ZCTA.

638 # Declarations

639 ## Ethics approval and consent to participate

640 Not applicable. This study analyzes only existing, publicly available, de-identified data,
641 analyzed at an aggregated scale.

642 ## Consent for publication

643 Not applicable

644 ## Availability of data and materials

645 The study was performed using publicly available data as described in the text. The datasets
646 used and analyzed during the current study are available from the original sources or from the
647 corresponding author on reasonable request.

648 ## Competing interests

649 The authors declare that they have no competing interests.

650 ## Funding

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This study was supported by a STAR Fellowship (Environmental Protection Agency FP917287) and a SAGE-IGERT traineeship (National Science Foundation Award Number 1144885) to BG. Publication fees were provided by SIUE. The funding bodies played no role in any aspect of the study (study design; data collection, analysis, or interpretation; or writing of the manuscript).

Authors' contributions

BG and JR conceived the study, developed study methods, and prepared the study data. BG performed the analyses and wrote the manuscript. JR and TM reviewed the paper and provided comments and edits. TM and BG addressed reviewer comments and revised the manuscript. All authors have read and approved the manuscript.

Acknowledgements

The data for CalEnviroScreen, which forms the basis of this analysis, has been made publicly available by California EPA's Office of Environmental Health Hazard Assessment. We thank Gregory S. Biging and Alejandra Benitez for helpful statistical feedback and John Balmes for constructively reviewing manuscript drafts. We thank Lynne Messer and Martin Lawrence for reviewing the manuscript.

References

1. DeFur PL, Evans GW, Cohen Hubal EA, Kyle AD, Morello-Frosch RA, Williams DR. Vulnerability as a function of individual and group resources in cumulative risk assessment. Environ Health Perspect. 2007;115:817–24. doi:10.1289/ehp.9332.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

671 2. Roth R, Barsi E. The “Community Need Index.” A new tool pinpoints health care disparities
672 in communities throughout the nation. Heal Prog. 2005;86:32–8.

673 3. Briggs D. Making a Difference: Indicators to Improve Children’s Environmental Health.
674 Geneva, Switzerland: World Health Organization Geneva; 2003.

675 4. National Research Council. Science and Decisions: Advancing Risk Assessment.
676 Washington, DC; 2009. http://www.nap.edu/catalog.php?record_id=12209.

677 5. Shankardass K, McConnell R, Jerrett M, Milam J, Richardson J, Berhane K. Parental stress
678 increases the effect of traffic-related air pollution on childhood asthma incidence. Proc Natl
679 Acad Sci U S A. 2009;106:12406–11. doi:10.1073/pnas.0812910106.

680 6. Solomon GM, Morello-Frosch R, Zeise L, Faust J. Cumulative environmental impacts:
681 science and policy to protect communities. Annu Rev Public Health. 2016;37:annurev-
682 publhealth-032315-021807. doi:10.1146/annurev-publhealth-032315-021807.

683 7. National Research Council. Exposure Science in the 21st Century: A Vision and a Strategy.
684 The National Academies Press; 2012. http://www.nap.edu/openbook.php?record_id=13507.

685 8. Scammell MK, Montague P, Raffensperger C. Tools for addressing cumulative impacts on
686 human health and the environment. Environ Justice. 2014;7:102–9.
687 doi:10.1089/env.2014.0016.

688 9. Lalloué B, Monnez J-M, Padilla C, Kihal W, Zmirou-Navier D, Deguen S. Data analysis
689 techniques: a tool for cumulative exposure assessment. J Expo Sci Environ Epidemiol.
690 2015;25:222–30. doi:10.1038/jes.2014.66.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

691 10. Messer LC, Jagai JS, Rappazzo KM, Lobdell DT. Construction of an environmental quality
692 index for public health research. Environ Heal. 2014;13:39. doi:10.1186/1476-069X-13-39.

693 11. Callahan MA, Sexton K. If cumulative risk assessment is the answer, what is the question?
694 Environ Health Perspect. 2007;115:799–806.

695 12. Sadd JL, Pastor M, Morello-Frosch R, Scoggins J, Jesdale B. Playing it safe: assessing
696 cumulative impact and social vulnerability through an environmental justice screening method
697 in the South Coast Air Basin, California. Int J Environ Res Public Health. 2011;8:1441–59.
698 doi:10.3390/ijerph8051441.

699 13. Meehan August L, Faust JB, Cushing L, Zeise L, Alexeeff G V. Methodological
700 considerations in screening for cumulative environmental health impacts: lessons learned from
701 a pilot study in California. Int J Environ Res Public Health. 2012;9:3069–84.
702 doi:10.3390/ijerph9093069.

703 14. Huang G, London J. Cumulative environmental vulnerability and environmental justice in
704 California's San Joaquin Valley. Int J Environ Res Public Health. 2012;9:1593–608.

705 15. Wilson S, Burwell-Naney K, Jiang C, Zhang H, Samantapudi A, Murray R, et al.
706 Assessment of sociodemographic and geographic disparities in cancer risk from air toxics in
707 South Carolina. Environ Res. 2015;140:562–8. doi:10.1016/j.envres.2015.05.016.

708 16. Shrestha R, Flacke J, Martinez J, Van Maarseveen M. Environmental health related socio-
709 spatial inequalities: Identifying "hotspots" of environmental burdens and social vulnerability.
710 Int J Environ Res Public Health. 2016;13:23 pp.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

711 17. Pearce JR, Richardson EA, Mitchell RJ, Shortt NK. Environmental justice and health: The
712 implications of the socio-spatial distribution of multiple environmental deprivation for health
713 inequalities in the United Kingdom. Trans Inst Br Geogr. 2010;35:522–39.

714 18. Alexeeff G, Mataka AY. CalEnviroScreen: A pathway to address environmental justice
715 issues in California. EM Mag. 2014; January:26–31.

716 19. Gaffron P, Niemeier D. School locations and traffic emissions — environmental (in)justice
717 findings using a new screening method. Int J Environ Res Public Health. 2015;12:2009–25.
718 doi:10.3390/ijerph120202009.

719 20. Los Angeles Collaborative for Environmental Health and Justice. Hidden Hazards. A Call
720 to Action for Healthy, Livable Communities. Los Angeles, CA; 2010.
721 https://www.libertyhill.org/news/reports/hidden-hazards-call-action-healthy-livable-
722 communities.

723 21. OEHHA. Responses to public comments on the method to identify disadvantaged
724 communities under Senate Bill 535. 2014.
725 http://www.calepa.ca.gov/EnvJustice/GHGInvest/Documents/SB535PubCom.pdf.

726 22. Morello-Frosch R, Pastor Jr. M, Sadd J, Porras C, Prichard M. Citizens, science, and data
727 judo. Leveraging secondary data analysis to build a community-academic collaborative for
728 environmental justice in southern California. In: Israel BA, Eng E, Schulz AJ, Parker EA,
729 editors. Methods in Community Based Participatory Research for Health. First Edit. San
730 Francisco, CA: John Wiley & Sons; 2005. p. 371–92.
731 http://www.wiley.com/WileyCDA/WileyTitle/productCd-0787980064.html.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

732 23. Faust J, August L, Bangia K, Galaviz V, Leichty J, Prasad S, et al. Update to the California
733 Communities Environmental Health Screening Tool, CalEnviroScreen 3.0. 2017.
734 https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-30.

735 24. Cushing L, Faust J, August LM, Cendak R, Wieland W, Alexeeff G. Racial/ethnic
736 disparities in cumulative environmental health impacts in California: evidence from a statewide
737 environmental justice screening tool (CalEnviroScreen 1.1). Am J Public Health. 2015;:e1–8.
738 doi:10.2105/AJPH.2015.302643.

739 25. Bierman P, Lewis M, Ostendorf B, Tanner J. A review of methods for analysing spatial and
740 temporal patterns in coastal water quality. Ecol Indic. 2011;11:103–14.

741 26. James FC, McCulloch CE. Multivariate analysis in ecology and systematics: panacea or
742 Pandora's box? Annu Rev Ecol Syst. 1990;:129–66.

743 27. Bivand RS, Pebesma EJ, Gomez-Rubio V. Applied Spatial Data Analysis with R. New
744 York, NY: Springer; 2008.

745 28. OEHHA. CalEnviroScreen Version 2.0. California Environmental Protection Agency
746 Office of Environmental Health Hazard Assessment (OEHHA). 2014.
747 https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-version-20. Accessed 13 Apr 2017.

748 29. OEHHA. Responses to major comments on the CalEnviroScreen 3.0 public review draft.
749 Sacramento, CA; 2017. https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-30.

750 30. Vieira VM, Villanueva C, Chang J, Ziogas A, Bristow RE. Impact of community
751 disadvantage and air pollution burden on geographic disparities of ovarian cancer survival in

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

752 California. Environ Res. 2017;156:388–93. doi:10.1016/j.envres.2017.03.057.

753 31. Committee on Models in the Regulatory Decision Process. Models in Environmental
754 Regulatory Decision Making. National Research Council, National Academies Press; 2007.

755 32. OEHHA. CalEnviroScreen 3.0. 2017.
756 https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-30. Accessed 7 Jun 2017.

757 33. OEHHA. New in CalEnviroScreen 3.0: changes since version 2.0. Sacramento, CA; 2017.

758 34. Babin SM, Burkom HS, Holtry RS, Tabernero NR, Stokes LD, Davies-Cole JO, et al.
759 Pediatric patient asthma-related emergency department visits and admissions in Washington,
760 DC, from 2001–2004, and associations with air quality, socio-economic status and age group.
761 Environ Heal. 2007;6:9. doi:10.1186/1476-069X-6-9.

762 35. Huang X, Lessner L, Carpenter DO. Exposure to persistent organic pollutants and
763 hypertensive disease. Environ Res. 2006;102:101–6. doi:10.1016/j.envres.2005.12.011.

764 36. Salomon JA, Haagsma JA, Davis A, de Noordhout CM, Polinder S, Havelaar AH, et al.
765 Disability weights for the Global Burden of Disease 2013 study. Lancet Glob Heal.
766 2015;3:e712–23.

767 37. Nord E. Disability weights in the Global Burden of Disease 2010: Unclear meaning and
768 overstatement of international agreement. Health Policy (New York). 2013;111:99–104.

769 38. Lessard LN, Alcala E, Capitman JA. Pollution, Poverty, and Potentially Preventable
770 Childhood Morbidity in Central California. J Pediatr. 2016;168:198–204.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

771 39. R Core Team. R: A language and environment for statistical computing. 2017.
772 http://www.r-project.org/.

773 40. Messer LC, Laraia BA, Kaufman JS, Eyster J, Holzman C, Culhane J, et al. The
774 development of a standardized neighborhood deprivation index. J Urban Heal. 2006;83:1041–
775 62.

776 41. Genberg BL, Gange SJ, Go VF, Celentano DD, Kirk GD, Latkin CA, et al. The effect of
777 neighborhood deprivation and residential relocation on long-term injection cessation among
778 injection drug users (IDUs) in Baltimore, Maryland. Addiction. 2011;106:1966–74.

779 42. Josse J, Husson F. missMDA□: A Package for Handling Missing Values in Multivariate
780 Data Analysis. J Stat Softw. 2016;70:1–31. doi:10.18637/jss.v070.i01.

781 43. Brunsdon C, Comber L. An Introduction to R for Spatial Analysis & Mapping. London,
782 UK: SAGE; 2015.

783 44. Anselin L. The Moran scatterplot as an ESDA tool to assess local instability in spatial
784 association. In: GISDATA Specialist Meeting on GIS and Spatial Analysis. Amsterdam, The
785 Netherlands: West Virginia University Regional Research Institute; 1993. p. 111–25.

786 45. Krivoruchko K. Spatial Statistical Data Analysis for GIS Users. Redlands, CA: ESRI Press;
787 2011.

788 46. Raftery AE. Bayesian model selection in social research. Sociol Methodol. 1995;25:111–
789 63.

790 47. Nagelkerke NJD. A note on a general definition of the coefficient of determination.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

791 Biometrika. 1991;78:691–2.

792 48. McCune B, Grace JB. Analysis of Ecological Communities. Gleneden Beach, OR: MJM
793 Software Design; 2002. www.pcord.com.

794 49. John JA, Draper NR. An alternative family of transformations. J R Stat Soc Ser C Appl
795 Stat. 1980;29:190–7. doi:10.2307/2986305.

796 50. Schielzeth H. Simple means to improve the interpretability of regression coefficients.
797 Methods Ecol Evol. 2010;1:103–13. doi:10.1111/j.2041-210X.2010.00012.x.

798 51. OEHHA. CalEnviroScreen 3.0 Draft: Public Comments. California Environmental
799 Protection Agency Office of Environmental Health Hazard Assessment (OEHHA). 2016.
800 http://oehha.ca.gov/calenviroscreen/comment/calenviroscreen-30-draft-public-comments.

801 52. Roberts EM, English PB, Grether JK, Windham GC, Somberg L, Wolff C. Maternal
802 residence near agricultural pesticide applications and autism spectrum disorders among
803 children in the California Central Valley. Environ Health Perspect. 2007;115:1482–9.

804 53. Alavanja MCR, Hoppin JA, Kamel F. Health effects of chronic pesticide exposure: cancer
805 and neurotoxicity. Annu Rev Public Health. 2004;25:155–97.

806 54. Marks AR, Harley K, Bradman A, Kogut K, Barr DB, Johnson C, et al. Organophosphate
807 pesticide exposure and attention in young Mexican-American children: The CHAMACOS
808 study. Environ Health Perspect. 2010;118:1768–74.

809 55. Balazs C, Morello-Frosch R, Hubbard A, Ray I. Social disparities in nitrate-contaminated
810 drinking water in California's San Joaquin Valley. Environ Health Perspect. 2011;119:1272–8.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

811 56. Weinhold I, Gurtner S. Understanding shortages of sufficient health care in rural areas.
812 Health Policy (New York). 2014;118:201–14. doi:10.1016/j.healthpol.2014.07.018.

813 57. Morgenstern H. Uses of ecological analyses in epidemiologic research. Am J Public Health.
814 1982;72:1336–44.

815 58. Thacker SB, Stroup DF, Carande-Kulis V, Marks JS, Rov K, Gerberding JL. Measuring the
816 Public's Health. Public Health Rep. 2006;121 February 2006:14–22.

817 59. Grubesic TH, Matisziw TC. On the use of ZIP codes and ZIP code tabulation areas
818 (ZCTAs) for the spatial analysis of epidemiological data. Int J Health Geogr. 2006;5:58.

819 60. Cushing L, Morello-Frosch R, Wander M, Pastor M. The haves, the have-nots, and the
820 health of everyone: the relationship between social inequality and environmental quality. Annu
821 Rev Public Health. 2015;36:193–209. doi:10.1146/annurev-publhealth-031914-122646.

822

823 **Figure Captions**

824 **Fig. 1.** Results of principal component analysis of CalEnviroScreen environmental hazard and
825 socioeconomic vulnerability variables across 7929 populated census tracts in California.
826 Variability explained by individual principal components is in parentheses. a. All variables PC1
827 versus PC2. b. All variables PC2 versus PC3. c. Environmental variables only. d.
828 Socioeconomic variables only.

bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

829 **Fig. 2.** Association between first principal component for all variables (PC1) and
830 CalEnviroScreen 3.0 score [23]. Each point represents a populated California census tract (n =
831 7929).

832 **Fig. 3.** Spatial pattern in first principal component of environmental variables (PC1Env),
833 aggregated at the zip code scale.

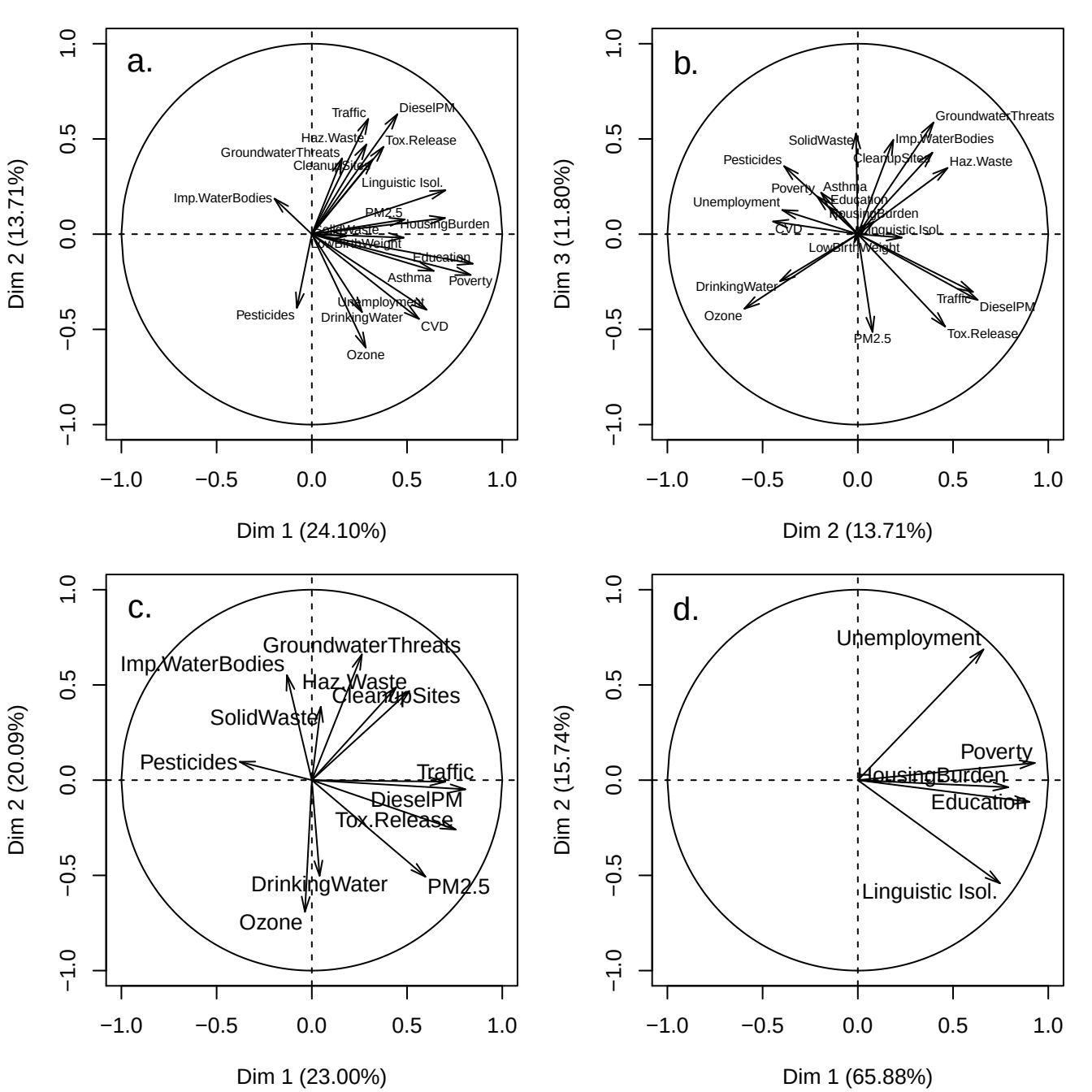
834 **Fig. 4.** Spatial pattern in first principal component of socioeconomic variables (PC1Soc), zip
835 code scale.

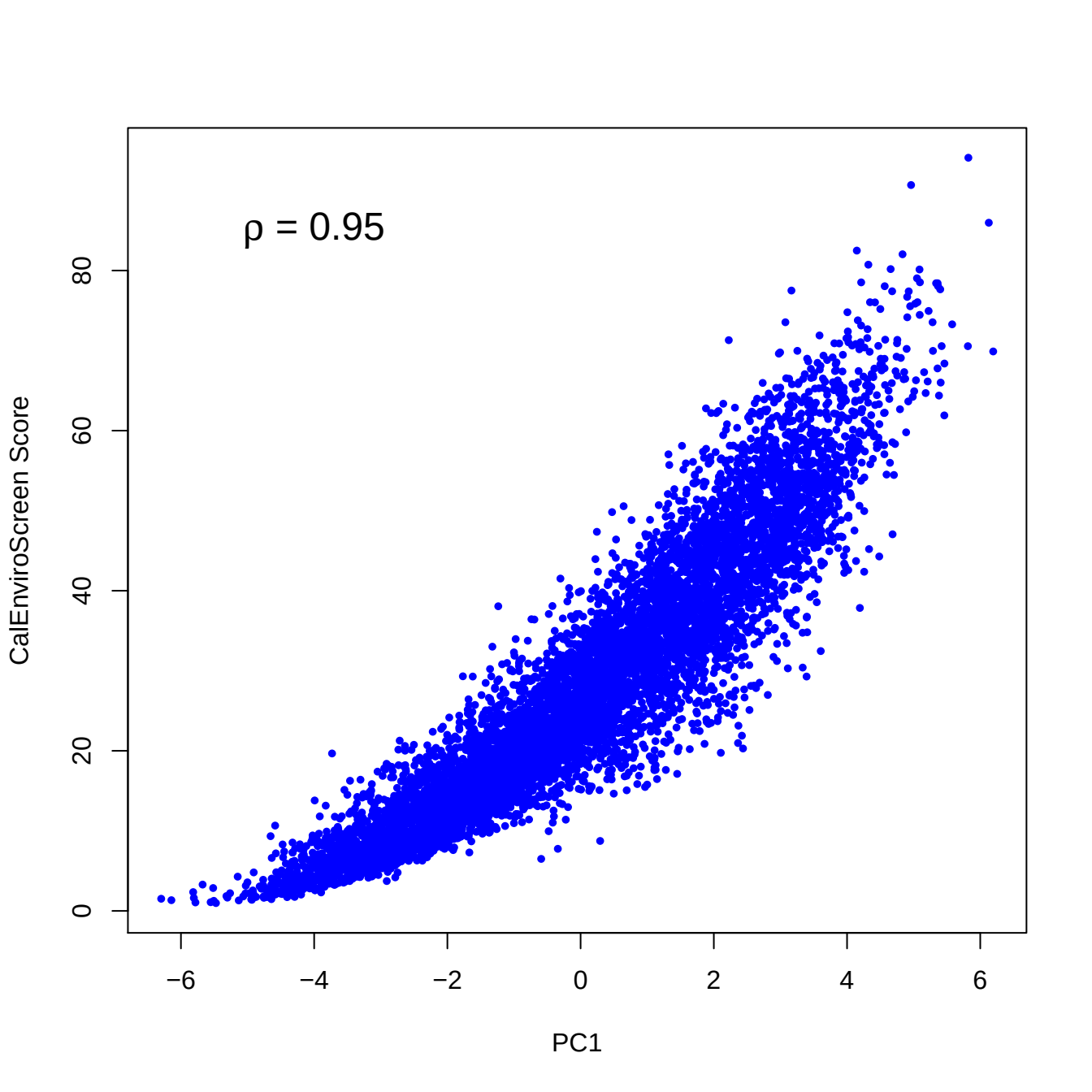
836 **Fig. 5.** Spatial pattern in CalEnviroScreen 3.0, zip code scale. These data have been previously
837 displayed at the census tract scale [23, 32], and are presented here to allow comparison to Figs.
838 3 and 4.

Table 1. Correlation matrix of underlying data from CalEnviroScreen 3.0. Data were transformed and analyzed using pairwise Pearson's correlation coefficients (r). Sample size ranged from 7694 to 8035 census tracts. Boldface indicates $|r| \geq 0.3$.

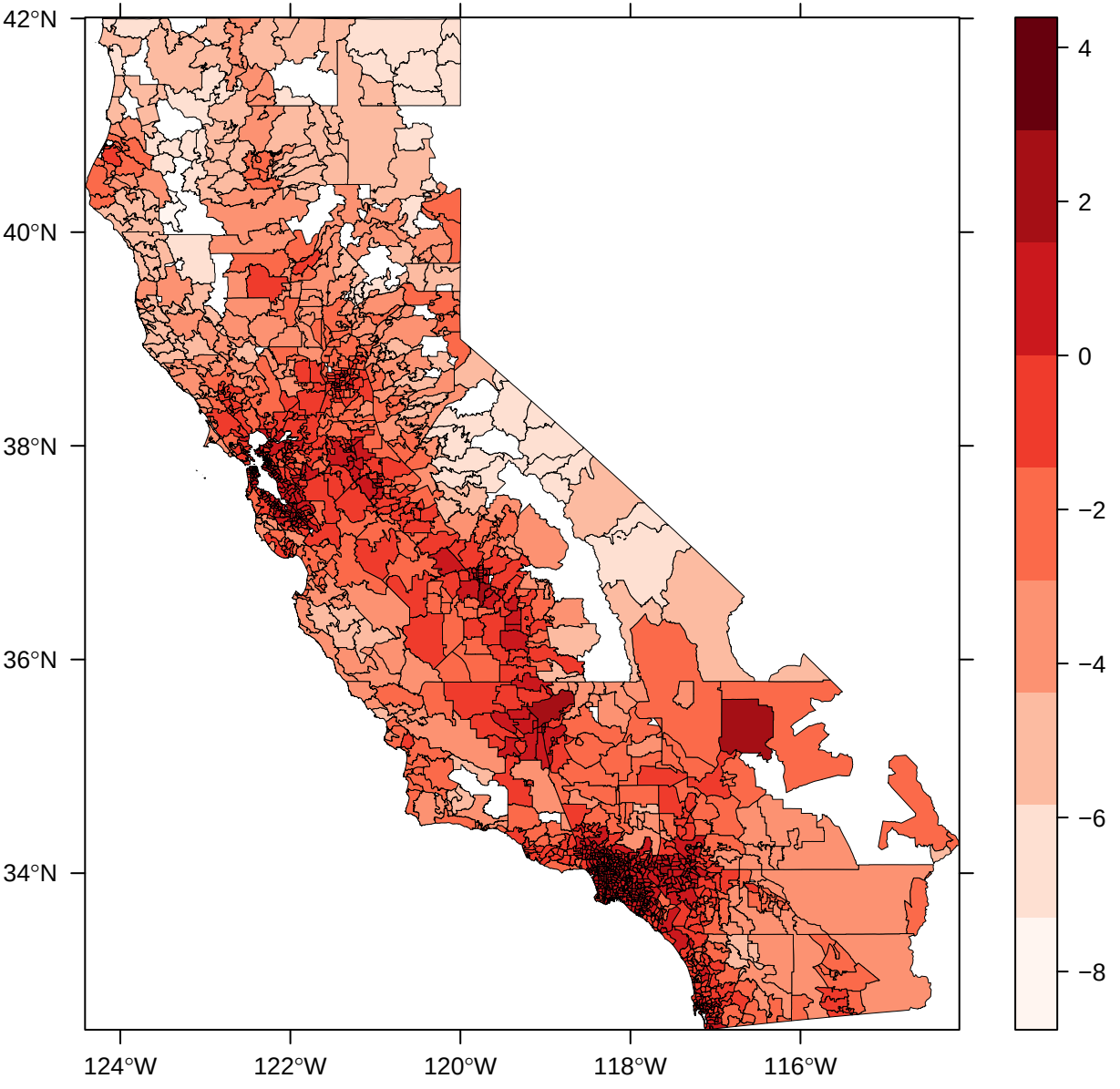
	Ozone	$PM_{2.5}$	Diesel PM	Traffic density	Drinking water	Pesticide use	Toxic release	Cleanup Sites	Groundwater threats	Hazardous waste sites	Water body impairments	Solid waste sites	Education	Linguistic isolation	Poverty	Unemployment	Housing burden	Asthma	Low birth rate
$PM_{2.5}$	**0.41**																		
Diesel PM	−0.16	**0.41**																	
Traffic density	−0.09	0.23	**0.56**																
Drinking water	**0.56**	**0.35**	−0.10	−0.02															
Pesticide use	0.08	−0.05	**−0.30**	**−0.30**	0.21														
Toxic release	0.06	**0.56**	**0.54**	**0.43**	0.09	−0.24													
Cleanup sites	−0.12	0.10	0.20	0.14	−0.01	−0.01	0.13												
Groundwater threats	**−0.33**	−0.05	0.10	0.10	−0.12	0.08	0.00	**0.44**											
Hazardous waste sites	−0.13	0.09	0.28	0.23	−0.03	−0.02	0.20	**0.49**	**0.39**										
Water body impairments	**−0.30**	−0.23	−0.09	−0.03	−0.17	0.14	−0.13	0.12	0.22	0.13									
Solid waste sites	0.03	−0.02	−0.17	−0.04	0.13	0.18	−0.02	**0.33**	0.29	**0.31**	0.17								
Education	0.06	−0.10	−0.18	−0.08	0.06	0.09	−0.12	−0.08	−0.03	−0.13	−0.12	0.19							
Linguistic isolation	−0.02	**0.30**	**0.43**	0.28	0.10	−0.07	0.31	0.22	0.13	0.18	−0.09	0.06	**0.70**						
Poverty	0.23	0.22	0.16	0.05	0.18	0.04	0.04	0.19	0.13	0.12	−0.11	0.17	**0.82**	**0.58**					
Unemployment	**0.30**	0.16	0.00	−0.05	0.19	0.04	−0.01	0.05	0.02	0.02	−0.08	0.10	**0.51**	0.24	**0.60**				
Housing burden	0.04	0.15	0.29	0.26	0.06	−0.16	0.17	0.15	0.10	0.14	−0.08	0.08	**0.57**	**0.50**	**0.72**	**0.40**			
Asthma	0.08	0.12	0.20	0.00	−0.04	−0.01	0.05	0.13	0.11	0.10	−0.05	0.08	**0.51**	0.25	**0.53**	**0.45**	**0.50**		
Low birth weight	0.10	0.18	0.19	0.11	0.06	−0.09	0.14	0.08	0.00	0.09	−0.09	0.03	**0.33**	0.25	**0.32**	0.27	**0.72**	0.28	
Cardiovascular disease	**0.39**	0.16	0.02	−0.08	0.18	0.08	0.03	0.03	−0.05	0.01	−0.14	0.08	**0.46**	0.17	**0.43**	**0.44**	0.21	**0.71**	0.24

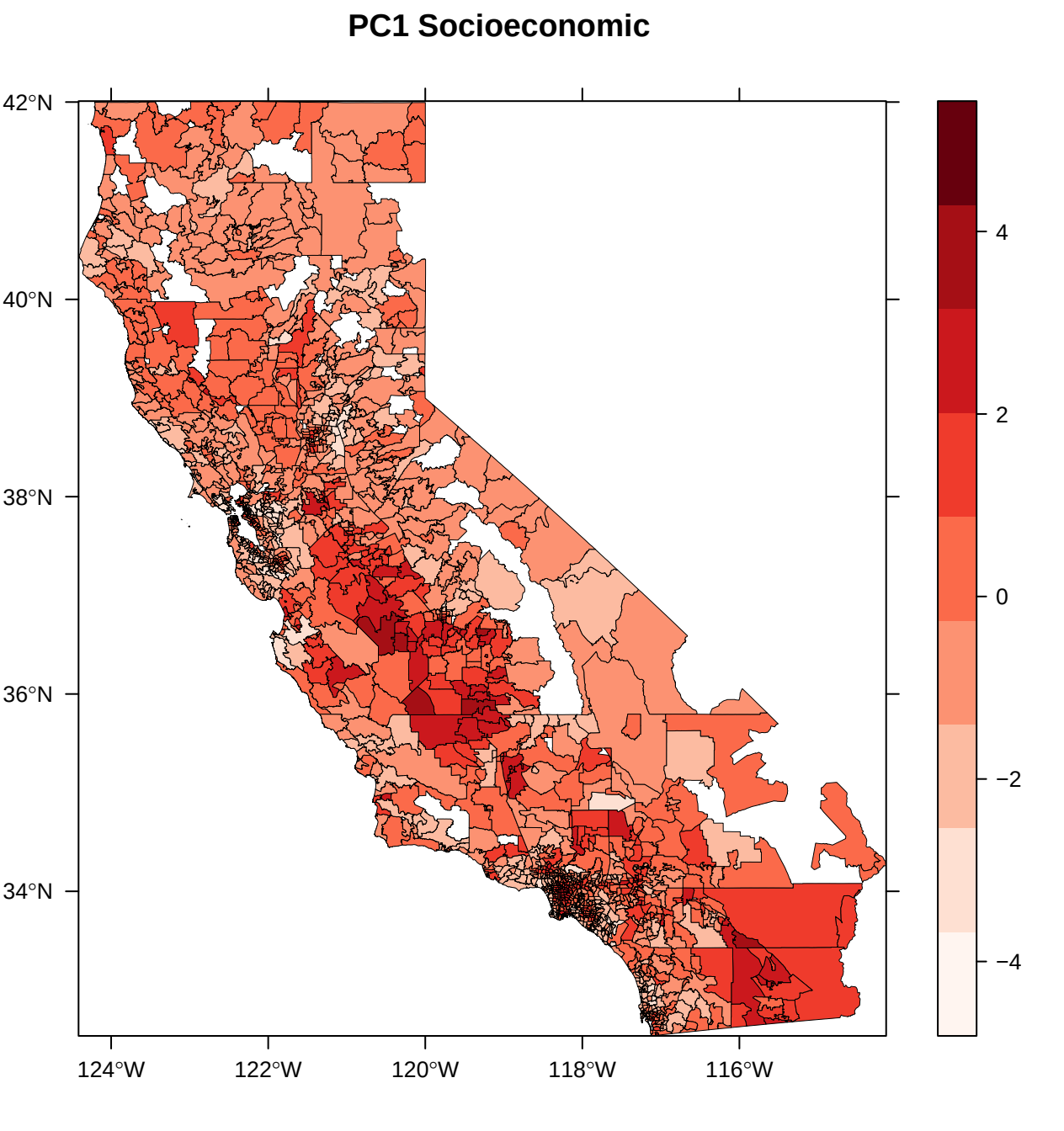
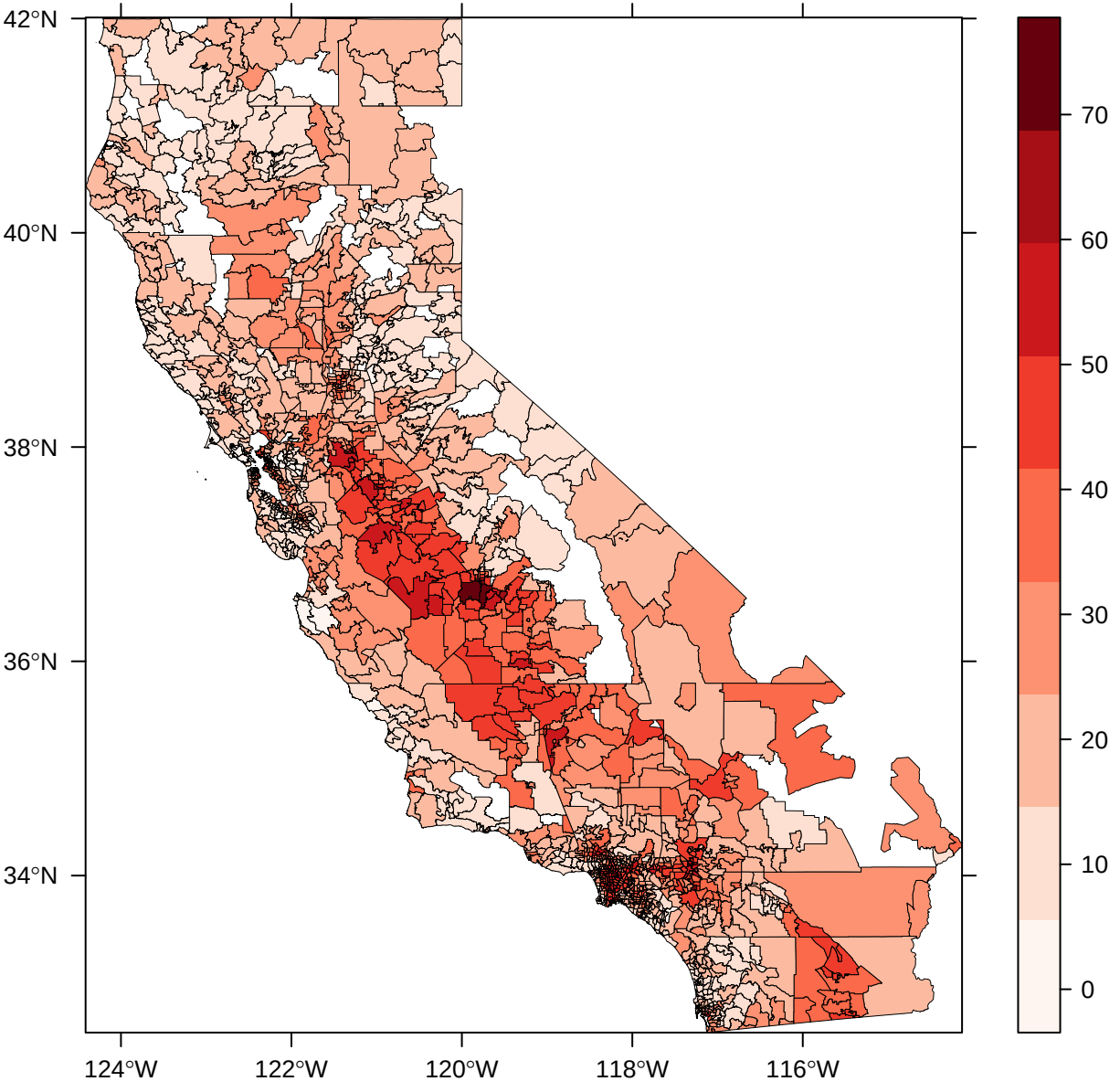
bioRxiv preprint doi: https://doi.org/10.1101/221739; this version posted November 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


a.
Traffic
DieselPM
Haz.Waste
Tox.Release
GroundwaterThreats
CleanupSites
Linguistic Isol.
Imp.WaterBodies
PM2.5
SolidWaste
HousingBurden
LowBirthWeight
Education
Asthma
Poverty
Unemployment
Pesticides
DrinkingWater
CVD
Ozone
Dim 1 (24.10%)
Dim 2 (13.71%)
b.
GroundwaterThreats
SolidWaste
Imp.WaterBodies
Pesticides
CleanupSites
Haz.Waste
Poverty
Asthma
Education
Unemployment
HousingBurden
CVD
Linguistic Isol.
LowBirthWeight
DrinkingWater
Traffic
DieselPM
Ozone
PM2.5
Tox.Release
Dim 2 (13.71%)
Dim 3 (11.80%)
c.
GroundwaterThreats
Imp.WaterBodies
Haz.Waste
CleanupSites
SolidWaste
Pesticides
Traffic
DieselPM
Tox.Release
DrinkingWater
PM2.5
Ozone
Dim 1 (23.00%)
Dim 2 (20.09%)
d.
Unemployment
Poverty
HousingBurden
Education
Linguistic Isol.
Dim 1 (65.88%)
Dim 2 (15.74%)


ρ = 0.95
CalEnviroScreen Score
0
20
40
60
80
−6
−4
−2
0
2
4
6
PC1

PC1 Environmental




PC1 Socioeconomic
42°N
40°N
38°N
36°N
34°N
124°W
122°W
120°W
118°W
116°W
4
2
0
−2
−4

CalEnviroScreen 3.0