pagepress

Quantifying altitude of human habitation in studies of human health using geographical name server data

Stephen Thielke,[1] Christopher G. Slatore,[2] William A. Banks[1]

[1]Puget Sound VA Medical Center, Geriatric Research, Education, and Clinical Centers, Seattle, WA; [2]VA Portland Health Care System, Center to Improve Veteran Involvement in Care, Portland, OR, USA

Abstract

Almost all studies examining the effects of altitude on human health have estimated the geographical altitude of defined regions, yet the primary interest lies in where people live, not the land around them. Populations are not homogenously distributed across altitudes. We propose a straightforward and computationally simple method for estimating the average altitude of habitation within the regional units for which health statistics are typically reported (such as counties). The United States Board on Geographical Names database contains records for over 2.7 million places, which can be processed to select places that are associated with human habitation. These points can easily be averaged by region yielding a representative altitude of

Correspondence: Stephen Thielke, Puget Sound VA Medical Center, Geriatric Research, Education, and Clinical Centers, 98108 Seattle, WA, USA.
Tel: +1.206.7642815 - Fax: +1.206.7642569.
E-mail: sthielke@u.washington.edu

human habitation within city, county, state regions, or by longitude and latitude zones. We provide an example of using this approach in a study of human health, and compare it with three other previously used methods of estimating altitude for counties.

Introduction

Recent studies have suggested that altitude may be associated with a variety of health outcomes, including better life expectancy (Ezzati *et al*., 2012), greater suicide (Haws *et al*., 2009; Helbich *et al*., 2013), more substance abuse (Kim *et al*., 2014), and higher dementia mortality (Thielke *et al*., 2015). Additional research of this type is likely to follow. Altitude may confound the relationship between health and other environmental or geospatial factors. All such research depends on reliable methods for determining altitude. While at first glance altitude seems to be an objective and fixed property of land regions, and thus easy to calculate, there are no authoritative methods for estimating it in studies of health. Analyses have used a variety of approaches, including comparing high and low altitude regions without quantifying them (Ishikawa *et al*., 2015); using the altitude of the highest point in the state or of the state capital (Haws *et al*., 2009); or (most commonly) computing the mean altitude of smaller land parcels within a state, sub-state, or county region (Huber *et al*., 2015; Kim *et al*., 2014; Selek, 2013). Almost all of these approaches share a methodological problem: they rely on *geological* rather than *human* features of the regions. They estimate a representative altitude for areas of land, but are not weighted for *where people live*. Presumably, the primary interest of studies of health and altitude is about altitude of habitation – where people spend their time – rather than of the land mass enclosed within regional boundaries. Because many health statistics are compiled by region such as county, it is important to estimate the mean altitude of human habitation in these regions. Even a cursory glance at maps of altitude and of population density shows that habitation is not homogeneously distributed across regions, both small and large. For instance, in Shasta county, California (CA), the vast majority of inhabitants live in one city (Redding), which lies at about 180 m above the mean sea level, but it is surrounded by high plains and mountains, which are sparsely populated. An altitude estimate based on geographical features would grossly overestimate the altitude of habitation.

In order to address this challenge, we sought to develop a straightforward method, which could be used to estimate the average altitude of human habitation within regions.

OPEN ACCESS

Materials and Methods

Possible methods

Estimating the representative mean altitude of human habitation within regions seems at first to be a relatively straightforward technical challenge. Region-specific estimates could account for population density, as by geocode. Smaller regions, for instance zoning improvement plan (ZIP) or telephone area code zones, could be weighted and aggregated in order to estimate the mean inhabited altitude of a county or state. Such an approach would assume that these smaller regions do not show considerable variations in altitude, which may not be true for regions that have a low population density (such as around mountains). More importantly, such analyses would depend on reliably identifying the mean altitude of smaller regions, which presents a computational challenge since their spatial boundaries are often complex. Linking data between various data sources increases the likelihood of errors, and using a large number of small units (such as over 40,000 ZIP codes for the United States, each with separate geological boundaries) can pose computational problems. Other high-resolution population-gridded datasets could be used to weight elevation data, but would have the same challenges of mapping geographical features onto regional boundaries.

Geographical name data

Because human habitation is associated with certain named features (*e.g.* school, airport) and not with others (*i.e.* gulch, peak), we decided to use the names of features to infer the degree of human habitation at different altitudes. These named features are also typically categorised by region, which makes it easy to summarise average altitude of habitation for different regions. Using place names from the United States Board on Geographical Names (GNIS) database, we developed a straightforward and computationally simple method for estimating the mean altitude of human habitation.

Data source

GNIS (http://geonames.usgs.gov) is the US national standard for geological nomenclature. It contains information about physical and cultural features. As of 2012, data from all 50 states and most foreign nations had been compiled. In 2014, there were about 2.7 million domestic names recorded and 9.8 million foreign names. The domestic database contains a separate entry for each feature, including feature identifier, feature name, feature class, State, County, county number, latitude, longitude, elevation in meters and elevation in feet. The features include various sites with geological or human significance, such as natural formations (*e.g.* lake, stream, cape, summit), man-made projects in natural settings (*e.g.* slough, canal, mine, tunnel) and civic spaces and buildings (*e.g.* populated place, locale, park, school, post office, building). The data fields are delimited by a special character making them easy to separate into arrays.

Method for generating average altitude of habitation

The method we propose assumes that population density roughly corresponds to the number of civic locations and buildings. Using string-matching functions, these can easily be selected from the feature class field. The fields related to habitation are *airport*, *building*, *city hall*, *civil*, *post office*, *library*, *locale*, *park*, *populated place*, and *post office* and *school*. The altitudes of each place thus defined can be compiled into an array, and averaged in order to estimate the altitude of habitation of the region in which they are contained.

We compared our estimates with other methods of measuring altitude that have been used in studies of human health: the altitude of the county seat in each county, the average altitudes of 1-km^2 land parcels and the highest point in the county. We examined the linear associations between our GNIS-based estimates and those of the three other approaches. We produced linear regression models, with the intercept set at zero. The regression coefficient represents the average proportional difference in a least-squares line between our estimates and those from the other methods. For instance, a coefficient of 1.4 would imply that the other method's estimates were 1.4 times higher on average that those using the GNIS data. The R^2 value is the coefficient of determination, a measure of the degree of linear (not absolute) association between the variables. We also tabulated the percentage of counties with a large relative difference (>50%) or absolute difference (>100 m), or both, compared to our estimates.

Example

A GNIS-based estimate of altitude of habitation was applied in a recent study examining the association between Alzheimer dementia mortality and altitude in CA counties (Thielke *et al*., 2015). Mortality rates were available for each of the 58 counties in the state, but there were no published estimates of the mean altitude of habitation for the counties. We used the approach above to identify place features associated with habitation. From the 121,684 place names in the GNIS dataset for CA, there were 4143 matches related to habitation, or about 71 per county. The elevation of all the locations in each county was averaged. The string matching and summarization were conducted using Perl scripts (https://www.perl.org/). We derived the altitudes of county seat, highest point in the state, and average altitude of land parcels from public sources.

Results

A scatterplot of the GNIS method compared to the three other methods is seen in Figure 1. The dashed line represents an exact correspondence between the variables.

The highest point in the county had almost no linear relationship with the GNIS-based estimates. Both county seat and average altitude of 1-km^2 land parcels showed a rough linear correlation, but the slopes were not close to 1.0, and there were many divergent estimates. As soon by the position of the slope line, the GNIS-based estimate was almost always lower than that of other methods.

Table 1 shows the regression coefficients, R-squared estimates, and percent of discordant cases.

The altitude of the county seat had the closest match with our estimate, and the slope of the line was 1.4, indicating that the county seat estimate was on average 1.4 times higher than that estimated using the GNIS data. About one quarter of the cases showed a significant difference from our estimates in either absolute or percentage difference. The average altitude of land parcels had a modest degree of correlation, and a slope of 2.1. About three-quarters of those estimates varied significantly from the GNIS-based estimates. The highest point in the county showed a very poor association for almost all counties, with a high regression coefficient (and almost no linear relationship, as seen in the graph), and with all of the cases varying greatly from the GNIS-based estimates.

Discussion

The use of GNIS data to determine altitude differs considerably from other ways of estimating a representative altitude for regions. In the absence of a *gold standard*, we believe that accounting for human habitation generates more accurate and relevant estimates than do techniques that rely only on geological features or a single point within the region (such as the county seat or the largest city). We propose that estimates which account for habitation patterns would be preferable as they have more utility for studies of human health than would geographically-based estimates.

Limitations

First, because there is no differentiation of place names based on their size, our technique may underestimate the population density of highly populous places. For instance, both a small-town airport and an international airport have a single entry, despite orders of magnitude difference in the numbers of people around them. It may be possible to use density maps to weight results further, although we do not know of a simple technique for doing so. Second, when comparing the GNIS-based method with other approaches, we did not examine other trends, such as nonlinear relationships. There was no evidence of nonlinear trends in the figure, except for the highest point in the county, which did not have a plausible association with altitude of human habitation.

Conclusions

This GNIS-based method for estimating altitude uses a publically available dataset and is easy to implement with any software that

Table 1. Comparison between estimates based on habitation and three other estimation methods.

|	Comparison estimation method		
	Highest point in county	County seat	Average of 1-km² land parcels
Regression coefficient (95% CI)	4.8 (3.6-5.9)	1.4 (1.3-1.6)	2.1 (1.8-2.4)
R²	0.32	0.86	0.80
Counties with difference of >50% with habitation-based estimate (%)	100	31	97
Counties with difference of >100 m with habitation-based estimate (%)	100	25	76
Counties with difference of both >50% and >100 m (%)	100	12	76

CI, confidence interval.
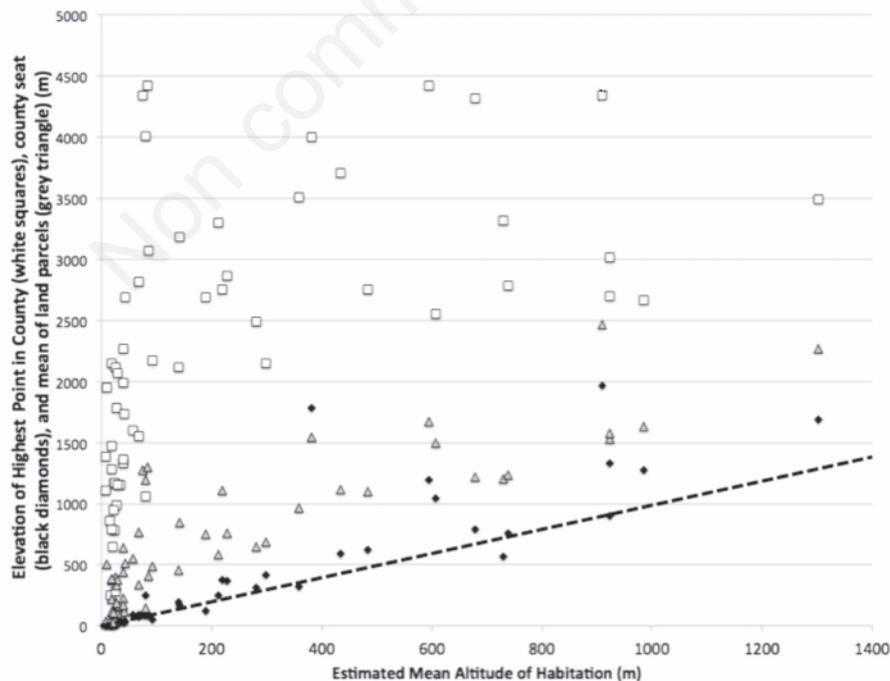
Figure 1. Scatterplot of estimated mean altitude of human habitation *vs* the altitude of the highest point in the county, the county seat and the average altitude of 1-km² land parcels for the 58 counties in California, USA. X-axis=human habitation. Dashed line represents a slope of 1.0 (*i.e.* an exact correspondence between the variables).

allows string matching. It can be linked to repositories of health data, which typically are aggregated by city, county or state. We anticipate that it can promote research about the effects of environmental factors on human disease. Mean latitude and longitude of different regions can also be estimated by the same approaches, easily allowing analyses of seasonal or diurnal patterns.

References

Ezzati M, Horwitz ME, Thomas DS, Friedman AB, Roach R, Clark T, Murray CJ, Honigman B, 2012. Altitude, life expectancy and mortality from ischaemic heart disease, stroke, COPD and cancers: national population-based analysis of US counties. J Epidemiol Commun H 66:e17.

Haws CA, Gray DD, Yurgelun-Todd DA, Moskos M, Meyer LJ, Renshaw PF, 2009. The possible effect of altitude on regional variation in suicide rates. Med Hypotheses 73:587-90.

Helbich M, Bluml V, Leitner M, Kapusta ND, 2013. Does altitude moderate the impact of lithium on suicide? A spatial analysis of Austria. Geospat Health 7:209-18.

Huber RS, Kim TS, Kim N, Kuykendall MD, Sherwood SN, Renshaw PF, Kondo DG, 2015. Association between altitude and regional variation of ADHD in youth. J Atten Disord (in press).

Ishikawa, M, Yamanaka G, Yamamoto N, Nakaoka T, Okumiya K, Matsubayashi K, Otsuka K, Sakura H, 2015. Depression and altitude: cross-sectional community-based study among elderly high-altitude residents in the Himalayan regions. Cult Med Psychiat 40:1-11.

Kim, TS, Kondo DG, Kim N, Renshaw PF, 2014. Altitude may contribute to regional variation in methamphetamine use in the United States: a population database study. Psychiatry Investig 11:430-6.

Selek S, 2013. Altitude, immigration and suicide rates: a study from Turkey. Psychiatry Investig 10:89-91.

Thielke S, Slatore CG, Banks WA, 2015. Association between Alzheimer dementia mortality rate and altitude in California counties. JAMA Psychiatry 72:1253-4.