Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Curr Res Ecol Soc Psychol. Author manuscript; available in PMC 2023 December 20.

Published in final edited form as:

Curr Res Ecol Soc Psychol. 2023 ; 5: . doi:10.1016/j.cresp.2023.100159.

Filling in the white space: Spatial interpolation with Gaussian processes and social media data

Salvatore Giorgi[a,*], **Johannes C. Eichstaedt**[b], **Daniel Preoţiuc-Pietro**[c], **Jacob R. Gardner**[a], **H. Andrew Schwartz**[d], **Lyle H. Ungar**[a,**]

[a]Department of Computer and Information Science, University of Pennsylvania, United States of America

[b]Department of Psychology & Institute for Human-Centered AI, Stanford University, United States of America

[c]Bloomberg, United States of America

[d]Department of Computer Science, Stony Brook University, United States of America

Abstract

Full national coverage below the state level is difficult to attain through survey-based data collection. Even the largest survey-based data collections, such as the CDC's Behavioral Risk Factor Surveillance System or the Gallup-Healthways Well-being Index (both with more than 300,000 responses p.a.) only allow for the estimation of annual averages for about 260 out of roughly U.S. 3,000 counties when a threshold of 300 responses per county is used. Using a relatively high threshold of 300 responses gives substantially higher convergent validity–higher correlations with health variables–than lower thresholds but covers a reduced and biased sample of the population. We present principled methods to interpolate spatial estimates and show that including large-scale geotagged social media data can increase interpolation accuracy. In this work, we focus on Gallup-reported life satisfaction, a widely-used measure of subjective well-being. We use Gaussian Processes (GP), a formal Bayesian model, to interpolate life satisfaction, which we optimally combine with estimates from low-count data. We interpolate over several spaces (geographic and socioeconomic) and extend these evaluations to the space created by variables encoding language frequencies of approximately 6 million geotagged Twitter users. We find that Twitter language use can serve as a rough aggregate measure of socioeconomic and cultural similarity, and improves upon estimates derived from a wide variety of socioeconomic, demographic, and geographic similarity measures. We show that applying Gaussian Processes to the limited Gallup data allows us to generate estimates for a much larger number of counties while maintaining the same level of convergent validity with external criteria (i.e., N = 1,133 vs. 2,954 counties). This work suggests that spatial coverage of psychological variables can be reliably extended through Bayesian techniques while maintaining out-of-sample prediction accuracy and that Twitter language adds important information about cultural similarity over and

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

*Corresponding author. sgiorgi@sas.upenn.edu (S. Giorgi). **Principal corresponding author. ungar@cis.upenn.edu (L.H. Ungar).

Declaration of competing interest

All authors have declared no conflicts of interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

above traditional socio-demographic and geographic similarity measures. Finally, to facilitate the adoption of these methods, we have also open-sourced an online tool that researchers can freely use to interpolate their data across geographies.

Keywords

Interpolation; Gaussian processes; Social media; Twitter; Geographical psychology

1. Introduction

Large geolocated data sets derived from psychological surveys or, recently, social media are an important tool for social scientific and public health research (Rentfrow, 2020; Hoover and Dehghani, 2020; Edo-Osagie et al., 2020). Such data sets have given further insight into personality, implicit racial attitudes, and subjective well-being, for example, by examining both their geographic variation and their relationships to other real-world outcomes (such as voting or policing) (Ebert et al., 2019; Hehman et al., 2019; Ward et al., 2021). In the case of geolocated social media data sets, community-level Twitter language has been used to predict health (Eichstaedt et al., 2015), behavior (Curtis et al., 2018), and psychological constructs (Giorgi et al., 2022b), in addition to standard socio-demographic and political outcomes (Culotta, 2014; Miranda Filho et al., 2015). The magnitude of these data sets (often containing millions of survey responses or billions of social media posts) allows researchers to study populations at multiple temporal and spatial levels, including both cross-national and sub-national levels (e.g., counties, cities, and neighborhoods) (Thomson et al., 2018; Cui et al., 2022; Bleidorn et al., 2016; Gibbons et al., 2019). These data sets are often less expensive and easier to collect (via online surveys or publicly available data streams) than those built from standard national polling techniques.

Despite the promise of large data sets that can be aggregated geographically, there are several methodological issues when doing fine-grained regional analyses, such as selection biases (i.e., non-representative samples of the underlying population (Giorgi et al., 2022a)) and limited geographic coverage due to data sparsity (Hoover and Dehghani, 2020). These sparsity issues are especially problematic when attempting to build stable estimates at (1) fine-grained spatial or temporal intervals (such as sub-state or sub-annual levels) and (2) low-population areas. Data sparsity issues can affect traditional survey and social media data sets alike. For example, the Centers for Disease Control (CDC) does not release mortality data for a U.S. county if the number of deaths is less than 10 (since a death record in this situation could reveal potentially private information). As a result of this, for example, when predicting maternal mortality with Twitter data, Abebe et al. (2020) were only able to work with outcome data from 197 out of roughly 3,000 U.S. counties, despite aggregating mortality over multiple years.

One standard approach when aggregating individual-level survey responses is to set a minimum threshold on the number of responses per spatial unit and ignore spatial units that do not meet this minimum. This approach is problematic in several ways, in addition to the fact that potentially useful data is discarded. First, there are no standards as to how to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

pick this minimum, and, thus, several minimums have been used across the literature (e.g., 50, 100, or 300) (Ebert et al., 2023; Matz and Gladstone, 2020; Stelter et al., 2022; Giorgi et al., 2018; Jaidka et al., 2020). As this threshold increases, the number of spatial units used in the final analysis decreases. Different choices of threshold can lead to coverage and results that are hard to compare between studies; for example, a 50 response minimum yielded 2,281 counties in one study (Ebert et al., 2023), while 1,208 counties met a 300 response minimum in another (Jaidka et al., 2020). Not only does the sample size decrease, but this decrease is non-random, typically removing rural counties, biasing the final sample towards urban areas with high population densities.

On the other hand, low minimum thresholds can produce unreliable spatial estimates. Giorgi et al. (2022b) showed that low minimum thresholds (< 500) resulted in low convergent validity between county-level language-based estimates of personality and self-reports. Similarly, Ward et al. (2021) showed low test-retest reliability for both county-level life satisfaction and happiness when using low minimum thresholds (< 200), with reliability stabilizing after 300 minimum responses. Thus, high minimum thresholds are needed to ensure the reliability of the spatial aggregates.

One possible solution to this trade-off between low thresholds (which retain data and help with representativeness) and high thresholds (needed for reliability) is to set a higher threshold and then interpolate across space using the more reliable estimates to “fill in the whitespace.” Several multivariate interpolation methods have been used successfully throughout geostatistics, such as inverse distance weighting and nearest neighbor interpolation (Sibson, 1981). Despite their success, many of these methods suffer from the fact that model parameters need to be manually selected and evaluated (e.g., in the nearest neighbor algorithm, one must select the number of neighbors a priori). Compounding this problem is the fact that neighbors may also be missing.

To address this problem, we propose Gaussian Processes (GP) to interpolate measures of interest from high-dimensional spatial, sociodemographic, and social media data. These methods are referred to as both kriging (Cressie, 1990), in spatial statistics, or Gaussian Process regression (Williams and Rasmussen, 2006), in machine learning. Gaussian Processes are uniquely equipped to deal with this problem by directly modeling the covariances between outcomes using a kernel function (also called a covariance function), which calculates the similarity or closeness between points. The kernel parameters, called length scales, model the extent to which a change in the inputs reflects changes in output. The length scales are *learned* (as opposed to being chosen a priori) from training data, which allows one to automatically identify the most predictive length scales for each feature (e.g., demographics or word topics). Additionally, the GP interpolations are probabilistic and, thus, produce empirical confidence intervals. These confidence intervals can then be used to optimally combine the interpolations with data that does not meet minimum thresholds, thus allowing researchers to maximize data use.

In this paper, we propose to address three research questions:

RQ1: What categories of community features are useful for interpolation?

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

RQ2: What is the minimum amount of data required for effective interpolation?

RQ3: Can supplemental data be used to improve interpolation accuracy?

For all three questions, we interpolate life satisfaction across U.S. counties, though we note that all methods are independent of the data used here. For **RQ1**, we note that traditional interpolation techniques (i.e., kriging) consider points close in 2- or 3-dimensional physical space. Here, we propose using higher-dimensional community characteristics such as demographics, socioeconomics, and social media language, in addition to standard geographic space. For **RQ3**, we propose to combine "missing" life satisfaction data (i.e., data from counties that do not meet our minimum count threshold) with the interpolated GP estimates to increase predictive performance. These estimates are optimally combined via inverse-variance weighting, using the uncertainty of the interpolations from the GP. Finally, we investigate the robustness of the interpolations by examining validity with external criteria. In order to make these methods available to the research community, we open-source a web interface for running interpolation over other data sources.[1]

2. Data

Our data falls into three classes: outcomes (measures which we want to interpolate), features (measures which we interpolate over, i.e., use to train a Gaussian process model), and external criteria (measures which we use to validate our interpolations). A total of 1,133 U.S counties had data available for all of the measures listed below. From this, we create train and test data sets using an 80%/20% split, which results in 905 counties for training and 228 for testing. The training data set is used to train the Gaussian Process model, whereas the test data set is used to evaluate the out-of-sample performance of the GP.[2]

2.1. Outcomes

Life satisfaction.—Life Satisfaction, an evaluative dimension of subjective well-being, is measured via psychometric self-reports using the Gallup-Sharecare Well-Being Index, a large national longitudinal survey. Participants are asked to respond to Cantril's ladder (Diener et al., 1999), which asks survey participants to evaluate their life as a whole: "Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?" We aggregate 2,035,511 responses from 2009 to 2016. To be included in the analysis, counties must have a minimum of 300 responses, an established minimum threshold (Ward et al., 2021; Jaidka et al., 2020).

2.2. Features

Categories of features are chosen in order to answer **RQ1**: which types of features are useful for interpolation above standard geography or location information (which are typically used when interpolating via Gaussian Processes). Features within each category are chosen to be

[1]https://county-interpolation.wwbp.org.
[2]Data and code available at https://osf.io/edjak/.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

representative of the category, and no single feature should be thought of as more important than another. In general, non-social media features are chosen for three reasons: (1) data is publicly available, (2) data is available for the majority of counties, (3) data is measured via the U.S. Census. Together, these three points allow these same variables to be used across similar interpolation problems, maximize spatial coverage, and limit biases.[3] The categories and features included here are by no means exhaustive and were intentionally selected to be general use in order to emphasize the utility of the methods, as opposed to optimizing on predicting life satisfaction.

Geography.—In order to compare counties close in physical space, we include latitude and longitude coordinates corresponding to the centroid of each county.

Demographics.—We include seven demographic variables, each collected from the U.S. Census American Community Survey (5-year estimates from 2010 to 2014): the percentage of the population living in a rural area, percentage of the population of Hispanic origin, population (logged to prevent skewness), median age, percentage of the population who identify as female, percentage married, and the percentage of African Americans living in the county.

Socioeconomics.—We include four socioeconomic variables which were, again, collected from the U.S. Census American Community Survey (5-year estimates from 2010 to 2014): median household income (logged to prevent skewness), percentage of the population with at least a Bachelor's degree, unemployment rate, and high school graduation rate.

Social media data.—We use the County Tweet Lexical Bank (Giorgi et al., 2018), a large open-source data set of U.S. county aggregated Twitter features. This data set is derived from a sample of 1.53 billion tweets from approximately 6 million Twitter users from 2009 to 2015. Each Twitter user is mapped to a U.S. county through self-reported location information available in the user's profile (e.g., "New York City native") or latitude/longitude coordinates associated with their tweets. Full details of the county mapping process can be found in Schwartz et al. (2013a). Each Twitter user must have at least 30 tweets in the data set, and each U.S. county needs at least 100 such users. In the end, a total of 2,041 counties met these thresholds. A set of 2,000 topics are extracted for each user and then averaged to the county level (across all users mapped to the county). Topics are automatically clustered groups of semantically related words and are created using Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a generative Bayesian topic model that assumes text documents are characterized by distributions over topics and topics are characterized by distributions over words. The specific set of 2,000 topics used in the current study was developed in previous work across a data set of 19 million Facebook posts (Schwartz et al., 2013b) and has been successfully used across several U.S. county-level studies (Jaidka et al., 2020; Giorgi et al., 2022b; Curtis et al., 2018). We ran Principal Component Analysis (PCA) across the 2,041 U.S. counties and created reduced feature sets

[3]While U.S. Census data may suffer from biases, such as non-response bias, it is typically considered a gold standard when doing spatial analysis.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of size 10, 15, 25, 50, and 100 principal components. This was done since the total number of topics (2,000) is larger than the number of observations (906 counties in the training data set). The difference in sizes between the observations and features could lead to overfitting, where the Gaussian Process model learns the training data too closely and, thus, will not generalize well to the unseen counties in the test data.

2.3. External criteria

These measures are chosen due to known associations with life satisfaction at both the individual level (Lee and Singh, 2020; Kahneman and Deaton, 2010; Wadsworth and Pendergast, 2014) and regional level (Arora et al., 2016; Lawless and Lucas, 2011). Similar to the feature variables, external criteria are chosen due to the fact that they can be robustly measured across most U.S. counties and, thus, there is ample data to compare against the interpolations. The external criteria are available for a total of 2,954 counties. Notably, this includes 1,821 counties not present in the train/test data since here we are interpolating life satisfaction across counties that do not have a gold standard. Finally, we note that this does not result in spatial coverage across 100% of the U.S., as some counties do not have publicly available data for all measures.

Life expectancy.—Life Expectancy is defined as the average number of years from birth that a person can be expected to live and is calculated using age-adjusted death rates from the population. Life expectancy is measured by the National Center for Health Statistics - Mortality Files from 2016–2018 and is obtained using the 2020 County Health Rankings (CHR) data.

Obesity.—Obesity is defined as the percentage of adults within a county that report a body mass index (BMI) of 30 or more. Data is reported from the 2013 Centers for Disease Control and Prevention (CDC) Diabetes Interactive Atlas and obtained from the 2017 County Health Rankings (Remington et al., 2015).

Income and education.—For both income and education, we use the measures listed above in the socioeconomic features: median household income (logged to prevent skewness) and percentage of the population with at least a Bachelor's degree. Both are collected from the U.S. Census American Community Survey (5-year estimates from 2010 to 2014). We note that both income and education are also used as features for interpolation. It may be the case (shown below) that interpolated outcomes correlate more with features than non-interpolated outcomes. Thus, using variables as both features and external criteria will artificially inflate associations. Therefore, we remove income and education from the feature data when comparing interpolations to external criteria.

3. Methods

3.1. Gaussian process regression

Gaussian Process (GP) regression is a machine learning algorithm, which is both supervised (i.e., learns a mapping between features x, such as sociodemographics, and labels y, such as life satisfaction) and probabilistic (i.e., model outputs can be used to determine uncertainty).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The full mathematical details are outside of the scope of the current study. However, the interested reader can consult Appendix A for more details on the kernel function or Schulz et al. (2018) for a full exposition on Gaussian Processes for psychology and social sciences.

At a high level, a Gaussian Process is defined by a mean and covariance function, also known as a kernel. This function induces similarity between pairs of data points. That is, given two data points x_i and x_j, if they are similar via the kernel, then their corresponding labels y_i and y_j will also be similar. For example, if the feature vectors x_i and x_j (e.g., socio-demographics and Twitter topics for counties i and j) are similar, then their corresponding life satisfaction values y_i and y_j will also be similar. Given the kernel function and training data set, we can fully specify the Gaussian Process model. Then, given a feature vector x^* from an unseen county (i.e., a county not included in the training data), we can estimate a life satisfaction score y^* by measuring the similarity between x^* and all points in the training data via the kernel.

Traditionally, Gaussian Processes are known as kriging in the field of geostatistics and have been used for decades to interpolate two- or three-dimensional spatial data, with applications in mining and environmental sciences (Chilès and Desassis, 2018). More recently, Gaussian Processes have been used in the field of Machine Learning (Williams and Rasmussen, 2006), incorporating methods and practices from deep learning. Using GPytorch (Gardner et al., 2018), a modern Python-based implementation of Gaussian Processes, we are able to *learn* the model hyperparameters (parameters which control the model learning) from the training data. Hyperparameters are typically chosen by searching over several potential values and evaluating the trained model at each value, which can be time-consuming and expensive. Learning hyperparameters allows non-specialists to train models using formal methods, thus extending these methods to a larger audience.

3.2. Inverse variance weighting

Inverse variance weighting is an optimal method for combining random variables via a weighted sum, such that the weighted average has the minimum variance across all possible weighted sums (Hartung et al., 2008). More formally, given a sequence of observations y_i with respective variances σ_i^2, the weighted average is computed as:

$$\hat{y} = \frac{\sum_i \frac{1}{\sigma_i^2} y_i}{\sum_i \frac{1}{\sigma_i^2}}. \tag{1}$$

We use inverse variance weighting to optimally combine the GP interpolations with the life satisfaction estimates (i.e., average of the person-level responses for each county) from counties that do not meet our minimum data threshold. We compute each county's c life satisfaction estimate y_c as

$$\hat{y}_c = \left(\frac{y_{GP,c}}{\sigma_{GP,c}^2} + \frac{y_{LC,c}}{\sigma_{LC,c}^2}\right) / \left(\frac{1}{\sigma_{GP,c}^2} + \frac{1}{\sigma_{LC,c}^2}\right), \tag{2}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where y_{GP} and σ^2_{GP} are the GP's life satisfaction estimate and variance, respectively, and y_{LC} and σ^2_{LC} are the life satisfaction estimate and variance from the low-count data, respectively. Specifically, inverse variance weighting is used to answer **RQ3** and is not used in either **RQ1** or **RQ2**.

3.3. Evaluation

We begin (**RQ1**) by considering which community features are useful for interpolation. Here, we abstract the standard notion of space (i.e., physical location) and consider counties neighbors if they are close across several non-geographic community characteristics: demographics, socioeconomics, and social media language use. We approachethis evaluation in two stages, by first considering non-language features (geography, demographics, and socioeconomics) and then adding social media language features on top of those. This was done for practical reasons: language data is not always available, while socio-demographics via the U.S. Census are readily available for most counties. Thus, we would like first to see how accurate our model is given available data and then how much of a boost we can get if we add in other data sources.

Next, for **RQ2**, we evaluate the minimum amount of data needed for effective interpolation. To do this, we randomly sample subsets of our training data. Specifically, we (1) randomly sample 10, 20, 40, 80, 160, 320, 640, and 905 counties from our training data set, (2) train a GP regression model, and (3) interpolate life satisfaction on our held-out test data set. This is repeated 50 times, and we report the average product-moment correlation across the 50 repetitions. We note that in Step (1), our complete training data set consists of 905 counties. Thus, we do not randomly sample at this stage, and this model is only evaluated once (as opposed to 50 times when randomly downsampling counties). We use the GPytorch package to implement the GP models (Gardner et al., 2018)[2], set a learning rate of 0.1, and iterate over the training data 500 times. Again, we first consider non-language features (spatial and socio-demographics features) and then add language on top of those.

For **RQ3** (can supplemental data be used to improve interpolation accuracy), we combine the GP interpolations with averages from data that does not meet our minimum count thresholds. We use the best-performing models from **RQ1** and considered the complete training data set (N = 905). This model is then used to both interpolate life satisfaction values $y_{GP,c}$ for each county c in our test data set, as well as estimate the variance $\sigma^2_{GP,c}$ for each county's interpolation. Next, using subsamples of the *participant-level* data (i.e., low-count data), we create life satisfaction estimates $y_{LC,c}$ for each county in the test data set. This is done by randomly sampling the participant-level life satisfaction responses and averaging these responses to the county level. Given that our minimum response threshold is 300 in the training data, we randomly sample sets of $n \in (25, 50, 100, 200)$ participants to produce the county averages. We then create new life satisfaction estimates $\hat{y}_c$ for county c using Equation (2), i.e., by optimally combining the GP interpolations with the low-count averages. We then correlate the final life satisfaction estimates $\bar{y}$ with the test data (gold standard life satisfaction values). This process was repeated 50 times, and the average product-moment correlation is reported.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Finally, we examine the external validity of the interpolated life satisfaction estimates from **RQ3**. First, we combine the training and test data (N=1133) above to train a GP model to interpolate life satisfaction from our non-language features. We do not use Twitter features in this analysis since they are unavailable for all counties and, thus, interpolations would have limited spatial coverage. We then interpolate life satisfaction across all counties that do not meet our minimum 300-participant response threshold (N=1,821). Next, we combine the GP interpolations with the low-count life satisfaction data. Finally, we correlate the combined estimates with external criteria: life expectancy, obesity, income, and education. We correlated gold standard life satisfaction with the external variables as a baseline. Here we show that the correlations using the complete data set (i.e., gold standard and interpolated life satisfaction across a larger sample of counties) are at least as accurate as the correlations using only gold standard life satisfaction (across a smaller sample of counties where all life satisfaction values are well-estimated). We note that two of the external criteria (income and education) are both used as features for interpolation. For these evaluations, we train the GP on all features except the variable used for external criteria so as to not bias the correlations.

4. Results

Table 1 shows the results for **RQ1**: which features are useful for interpolation. All GPs use a radial basis function (RBF) kernel function with a single length parameter for all features. See Table D.7 for other kernels functions (linear and multi-length RBF). Here we see that the geographic features (latitude and longitude) have the lowest out-of-sample accuracy. Despite this, adding geographic features to either socioeconomics or demographics gives a substantial boost (r = 0.57 and 0.57, respectively) over either socioeconomics or demographics alone (r = 0.46 and 0.49, respectively). Similarly, a GP trained on all non-language features out performs all subsets (r = 0.65). Finally, when combining Twitter with the non-language variables we see a boost in performance above all non-language alone, which is maximal when using 25 Twitter PCA components (r = 0.70).

Results for **RQ2** (what is the minimum amount of data required for effective interpolation) are in Fig. 2. Again, all GPs used a RBF kernel with a single length parameter learned across all features. Results show a mostly monotonic increase (i.e., within the confidence intervals) in accuracy as the training size increases.

In Fig. 2(b) we see the results of combining the language features (i.e., PCA reductions of the LDA topic space) with the non-language features (i.e., latitude/longitude, demographics, and socioeconomics). Here we see a slight improvement when adding in language, with the final predictive accuracy using the entire training data with 25 PCA components resulting in a product moment correlation of 0.70. These models (25 Twitter PCA components + all non-language variables) are then used in all subsequent analyses, except comparing against external criteria, where Twitter language is dropped to maximize spatial coverage. We also see that the smallest PCA components set (10 and 15) results in the highest accuracy when training on smaller sample sizes, outperforming the non-language variables at 320 samples.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2 shows the results of combining the GP interpolation values with the average life satisfaction scores as calculated from random samples and combined via inverse variance weighting (**RQ3**). The first row *GP Interpolation* is the results from the model shown in Table 1 using all non-language features. The *GP Interpolation* row remains constant since it does not depend on maximum number of samples used in averaging life satisfaction scores (i.e., this ignores the low-count data). The *Average Life Sat.* row is simply the county average life satisfaction score using random samples of 25, 50, 100, and 200 participants. The *Combined* row optimally combines (via inverse variance weighting) the data from the previous two rows: the GP interpolation and the average low-count estimate.

Finally, we validate the interpolated life satisfaction against external criteria: life expectancy, obesity, income, and education. Here our baseline is the correlation with the non-interpolated gold standard average life satisfaction (i.e., at least 300 participant responses per county). We also consider a simple "state average" interpolation baseline, where we assign the state-average life satisfaction score to all counties which do not meet the minimum data threshold. Results are in Table 3. The *state average* and *average life satisfaction* have the lowest correlation across all four external variables. We see that the combined model shows similar effect sizes as baseline but includes roughly 3,000 observations, as opposed to the 1,133 observations in the baseline model. We ran a bootstrapping test, to assess statistically significant differences between the correlation across the high-count counties and the combined model (interpolations plus low-count data). Here we randomly sample (with replacement) 1,133 counties and correlate the life satisfaction measure with the external criteria, subtracting the two correlations. This process is repeated 10,000 times, and no significance is found if the number 0 is within the 95% confidence interval on the difference in correlations. Across all four external criteria we found no difference in effect size between the gold standard and the interpolated correlations. Fig. 1 shows the geographic distribution of the 1,133 gold standard counties as compared to the 2,954 counties with either gold standard or interpolated life satisfaction.

5. Discussion

This study found that Gaussian Processes, formal Baysian models, can be used to empirically interpolate life satisfaction across U.S. counties using high-dimensional community characteristics, including social media language. Furthermore, GPs provide confidence intervals for the interpolations, which allows us to optimally combine the interpolations with estimates from noisy data, data typically discarded using standard thresholding methods. Finally, these methods allow us to estimate life satisfaction across most of the U.S., as seen in Fig. 1.

The results for **RQ1** (which community characteristics make an accurate interpolation space), as seen in Table 1, show that each set of variables contributes uniquely to the overall accuracy of the model. That is, when combining sets of feature (e.g., geographic and socioeconomic), we see an increase in accuracy above each feature set alone. Across all models, we see that including geographic features increases predictive accuracy despite geography alone being the least accurate model. While this is not surprising, since adjacent counties are often similar, it suggests that both (1) geographic proximity should not be

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ignored when considering high-dimensional interpolation spaces and (2) adjacency alone does not fully capture the geographic variation of life satisfaction.

For the Twitter experiments in Table 1, we see similar performance across each set of PCA components. We also see a boost in performance when combining both the Twitter and non-language feature sets, with an accuracy of 0.70 (product-moment correlation) for the best performing model (25 PCA components + All non-language). This is a statistically significant 7.69% increase over "All non-language" alone, suggesting that language from social media is capturing predictive signal not presents in standard geographic or socio-demographic indicators.

While all Twitter experiments produced equivalent predictive accuracies in Table 1, we see a different story when examining how much data is needed to create an accurate interpolation space (**RQ2**). In Fig. 2, we see smaller feature spaces (10 and 15 PCA components) resulting in higher predictive accuracy at smaller training sizes. This is most likely due to the GP over-fitting on the training data and not generalizing on unseen data since the size of the feature spaces (25, 50, and 100 plus the 13 non-language features) is larger than the number of training observations (e.g., 10, 20, and 40 counties). Thus, while accurate interpolations are attainable from a small number of observations, one must consider the dimensionality of the interpolation space in reference to the number of observations.

For both the language and non-language feature spaces in Fig. 2, predictive accuracy tends to stabilize around 320 observations (except for the 100 Twitter PCA components). This suggests that data from 300–400 counties may be sufficient to accurately interpolate life satisfaction across the U.S. and, thus, may be used as a minimum sample size in future data collection efforts (**RQ2**). Previous studies have shown that effect sizes tend to vary (in both direction and magnitude) according to the construct at both the individual and regional level (Elleman et al., 2020; Giorgi et al., 2022b; Eichstaedt et al., 2021) and, thus, we do not wish to over-generalize the claim.

Finally, we answer **RQ3** in the affirmative: we can use data from counties that do not meet the minimum thresholds. Table 2 shows that the optimal combination of GP interpolations and estimates from low-count data provides more accurate estimates than from the GP interpolations alone, thus lowering the minimum data thresholds. In Table 3, we also see that we can interpolate life satisfaction across the entire U.S. without a considerable reduction in correlation with external variables. While we see a reduction in effect sizes, the sample size increases from 1,113 to 2,954 counties. Leveraging the low-count data allows us to represent more of the U.S. population and, importantly, a section of the population that tends to be discarded due to data quality issues (i.e., sparsely populated rural areas).

Importantly, the methods introduced here estimate missing data at the spatial level and are agnostic to how those spatial level values are aggregated from person-level responses. The U.S. county-level life satisfaction values used throughout the paper are simple averages of person-level data. However, one could use more sophisticated aggregation methods such as multilevel regression with poststratification, which could help mitigate selection and non-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

response biases common when dealing with spatially aggregated person-level data (Hoover and Dehghani, 2020).

As seen by the small number of white counties in Fig. 1(b), it is not always possible to interpolate over the entirety of the U.S. This is due to the fact that secondary data must be available to interpolate over. Such data is not always available and is highly dependent on the type of data and the spatial level. For example, mortality data from the CDC is not always available (for privacy reasons) and can become even more sparse at the sub-county level (e.g., Census tracts or Census blocks). Thus, using mortality data as a feature to interpolate over may be limiting. Alternatively, demographic variables from the U.S. Census are generally available at smaller spatial resolutions and may be useful across many types of interpolation tasks.

Before using these methods, one should consider *why* data is missing. On the one hand, these methods allow researchers to estimate measures across populations that are typically ignored or excluded. At the same time, these methods have ethical and privacy concerns. For example, as discussed above, the CDC does not release mortality data for spatial units with less than ten deaths due to privacy reasons. Therefore, interpolating such measures could open up the risk of exposing individuals. Similarly, governments and private companies could use such methods to track protected or private measures across communities without their consent. In addition to these privacy concerns, data may be missing for other non-random reasons. For example, in the current study, the data used to train the Gaussian Process are collected from mostly urban areas and then used to interpolate across rural areas. This may bias the interpolations in that the Gaussian Process can only learn the relationship between the features and Life Satisfaction in the context of urban areas. We have attempted to mitigate this bias by using the percentage of the population living in a rural area as a feature, though there is no reason to believe the trained model can fully generalize from urban to rural contexts without including more rural areas in the training process.

This study is limited in several ways. First, the methods were evaluated using a single construct: life satisfaction. While the proposed methods are more general and not tied to particular data sources or outcome measures, there may be varying performance depending on the construct to be interpolated. Second, each county's linguistic representation is measured via LDA topics. While LDA is used extensively throughout natural language processing and computational social science (as well as in other geographic studies), there exist many other ways to measure language, including other topic models (such as Latent Semantic Analysis (Landauer et al., 1998) and BERTopic (Grootendorst, 2022)) as well as more modern contextual embeddings, such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020). While contextual embeddings have become standard across many computational tasks (Rogers et al., 2021), they have yet to be evaluated in the context of spatial language or community-level prediction tasks. We also note that this paper's social media data set consists of over 1.5 billion tweets. Running models such as BERT over such a large data set is computationally expensive (i.e., must be run over each tweet) when compared to extracting LDA topics, which is linear and can thus be done on aggregated county representations. Finally, there are better ways to measure spatial proximity in the U.S. than the geographic features (latitude and longitude coordinates of county centroids).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For example, the average distance between two adjacent counties on the East Coast (where counties are smaller on average) is smaller than the average two adjacent counties on the West Coast (where counties are much larger). Thus, the GP kernel's length parameter will perform differently across the entire U.S. To properly handle this issue, one could define a GP over a graph, where adjacent counties are connected nodes within the graph (Ng et al., 2018), thus circumventing the need for an invariant distance measure.

One must also consider the downstream applications of the interpolations. For example, Table B.4 shows the correlations between the socio-demographic features used to create the interpolation space and life satisfaction, both the gold standard estimates (i.e., the test data set) and the interpolations. Here, the interpolated life satisfaction correlates more with the socio-demographic features than the gold standard. Thus, one may want to avoid using the socio-demographic features in any downstream applications as the results may be confounded. Similarly, these increases in association with features could amplify biases in the data set. For example, Table B.4 shows an increase in the association between population density and life satisfaction. Given that low population areas were mostly excluded from the data set, this increase may not reflect the true relationship between well-being and population density and, furthermore, may have changed in the wrong direction.

Recommendations.

Choosing a feature set to interpolate over should be done on a case-by-case basis with an end task in mind: What will the interpolated values be used for? Using life satisfaction as an example, one could make a case for interpolating over income since income and life satisfaction are highly correlated and, thus, income could be considered a proxy for life satisfaction. On the other hand, the resulting interpolated life satisfaction values may be indistinguishable from income. If this end task is to study relationships between income and (interpolated) life satisfaction, then the results will be highly confounded. At the other extreme, one *must* select features that correlate with the outcome of interest for the Gaussian Process to learn how to interpolate. Thus, confounds may be unavoidable. One may also consider selecting features on which the missing spatial units are biased. In the present study, we selected the percentage of the population living in a rural area since the counties where no life satisfaction data was available were highly rural. This was done under the assumption that the Gaussian Process could learn the relationship between life satisfaction and urban/rural counties and thus model this when interpolating over the highly rural areas. Due to data constraints, we were unable to fully explore this. In the end, we recommend three rules when choosing a feature set. First, a feature should be excluded if the interpolated values will be used to study this feature (i.e., the end task). Second, while not highly accurate on their own, simple latitude and longitude coordinates increase accuracy when combined with other features and are not immediately confounding. Third, if there is reason to believe a feature is confounding downstream results, one should remove that feature from the interpolation space and see if downstream results still hold. Finally, we highly recommend transparency when reporting interpolated results: feature sets should be reported and possible confounds highlighted.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

6. Conclusions

Gaussian Process regressions can accurately interpolate U.S. county-level life satisfaction using spatial proximity, socio-demographics, and social media language. Importantly, these methods allow for principled estimation, where model parameters are empirically learned from training data, as opposed to chosen a priori. The interpolations can be optimally combined with sparse data from under-sampled counties to produce accurate and valid life satisfaction estimates for the majority of counties in the U.S. By utilizing data from under-sampled counties, larger sections of the population are present in the final data set, leading to potentially less biased and more representative spatial aggregates.

Acknowledgements

This work was supported in part by NIH Grant Number R01 MH125702-03 (Smart and Connected Health: Improving the Robustness of Monitoring Mental Health of Populations from Social Media) and by Stanford's Institute for Human-Centered AI (to JCE).

Data availability

Data is available on OSF: https://osf.io/edjak/.

Appendix

Appendix A. Details on Gaussian process regression kernels

The kernel κ induces similarity between pairs of data points evaluated at f: given any two vectors x_i, x_j, if these vectors are similar via κ then $f(x_i)$ and $f(x_j)$ will also be similar. We use a squared exponential or Radial Basis Function as our kernel, which defines a smooth function between neighboring points:

$$\kappa(x_i, x_j) = \exp\left(\frac{-(x_i - x_j)^T (x_i - x_j)}{2l^2}\right). \tag{A.1}$$

Here l is the lengthscale parameter which measures the rate of change for each feature in the training data (e.g., a larger lengthscale corresponds to smaller change). One can either use a single lengthscale for each feature in the training data or use different lengthscales for each feature. The lengthscale is traditionally a tuning parameter (e.g., selected through a grid search). Using GPytorch (Gardner et al., 2018), we are able to learn the lengthscale l from the training data.

Appendix B.: Correlations with socio-demographics

In Table B.4 we show the correlations between both the known life satisfaction estimates and the interpolated values and the sociodemographics variables used to create the interpolation space. Across all socio-demographics variables we see a larger correlation with the life satisfaction interpolations than the gold standard life satisfaction.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table B.4

Product moment correlations between socio-demographics and life satisfaction (228 counties in the test data set), using both the average life satisfaction from high-count counties and the interpolated estimates. Here we use the best performing model from Fig. 2(b): 25 Twitter PCA components + all non-language features.

	Variable	Average life satisfaction (high-count counties)	Interpolated life satisfaction
Demographics	% Rural	−0.31 [−0.42, −0.18]	−0.45 [−0.55, −0.34]
	% Hispanic	0.07 [−0.06, 0.20]	0.18 [0.05, 0.30]
	Population	0.17 [0.04, 0.29]	0.31 [0.18, 0.42]
	Median Age	−0.14 [−0.26, −0.01]	−0.29 [−0.40, −0.16]
	% Female	0.11 [−0.02, 0.24]	0.12 [−0.01, 0.25]
	% Married	0.17 [0.04, 0.30]	0.10 [−0.03, 0.23]
	% Africa American	0.11 [−0.02, 0.24]	0.17 [0.04, 0.29]
Socioeconomics	Median Household Income	0.41 [0.30, 0.51]	0.61 [0.53, 0.69]
	% Bachelor's degree	0.31 [0.19, 0.42]	0.53 [0.43, 0.61]
	Unemployment Rate	−0.41 [−0.51, −0.30]	−0.47 [−0.57, −0.37]
	High School Graduation Rate	0.13 [−0.00, 0.25]	0.06 [−0.07, 0.19]

Appendix C.: Spatial autocorrelation

Here we calculate the spatial autocorrelation of all of the non-language features used to train the Gaussian Process. Spatial autocorrelation measures the degree to which counties closer in space have more similar feature patterns than more distant counties. We use Moran's I to measure spatial autocorrelation (Moran, 1950), which ranges from −1 (dispersion or clustering of dissimilar values) to 1 (clustering of similar values), where 0 represents randomness. This is done to examine if the Gaussian Process can learn better interpolations from features which have more spatial autocorrelation (i.e., situations where adjacency better captures geographic variation).

Results are in Table C.5. Here we see higher Moran's I, on average, across the socioeconomic features (Moran's I = 0.44) versus the demographics (Moran's I = 0.40). This may support the hypothesis that Gaussian Processes can learn better interpolations when the adjacency in features captures geographic variation, since the socioeconomic features had higher prediction accuracy than demographics for small training sizes (see Fig. 2). We also see that Percent Hispanic has the highest spatial autocorrelation and Percent Married has the lowest.

One hypothesis is that the degree of spatial autocorrelation could be driving the change in correlations found in Table B.4. To test this, we correlate the percentage increase with the Moran's I values, but see no significant relationship (product-moment correlation of −0.50, $p = 0.12$).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Similarly, we calculate Moran's I for the interpolated life satisfaction values to see how interpolation effects spatial autocorrelation. We note that this is done on the test data set features, unlike the analysis above which considers the features across the training data set. Results are shown in Table C.6. Moran's I in the test data set is 0.04, showing that the life satisfaction scores (from the high-count counties) have little spatial autocorrelation. The results show that life satisfaction interpolations (regardless of the feature set) have higher spatial autocorrelation than the non-interpolated life satisfaction. Looking across specific feature sets, we see the geography features producing interpolations with the highest Moran's I (0.82). This is expected since spatial autocorrelation is measured via adjacency and the geography features are the only features which measure physical space. Notably, adding the geography features to other feature sets (socioeconomics or demographics) decreases Moran's I. We also see increases in Moran's I as we increase the number of Twitter PCA components.

Table C.5

Moran's I (a measure of spatial autocorrelation) for each feature in the test data set (905 counties). All results significant at $p < 0.01$.

	Variable	Moran's I
Demographics	% Rural	0.39
	% Hispanic	0.78
	Population	0.41
	Median Age	0.35
	% Female	0.23
	% Married	0.10
	% Africa American	0.55
Socioeconomics	Median Household Income	0.55
	% Bachelor's degree	0.35
	Unemployment Rate	0.48
	High School Graduation Rate	0.40

Table C.6

Spatial autocorrelation (Moran's I) of interpolated life satisfaction values across the test set data (228 counties). Moran's I for the true values (average life satisfaction from high-count counties) in the test set 0.04.

	Moran's I
Geography	0.82**
Socioeconomics	0.32**
Socioeconomics+ Geography	0.24*
Demographics	0.34**
Demographics+ Geography	0.24*
All non-language	0.24*
Twitter, 10 PCA components	0.18

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

	Moran's I
+ All non-language	0.29**
Twitter, 15 PCA components	0.27**
+ All non-language	0.27**
Twitter, 25 PCA components	0.33**
+ All non-language	0.28*
Twitter, 50 PCA components	0.34**
+ All non-language	0.29**
Twitter, 100 PCA components	0.37**
+ All non-language	0.30**

** $p < 0.01$,
* $p < 0.05$.

Appendix D.: Kernels

In Table D.7 we investigate the effect of using both linear and RBF kernels. As opposed to the results in Table 1, the RBF kernel here learns a separate length parameter for all variables in the model (e.g., it learns 13 length parameters for the "All non-language" model) as opposed to learning a single length parameter used across each of the 13 variables.

Table D.7

Out-of-sample prediction accuracy (product moment correlations with 95% confidence intervals).

	Number of Features	Linear	RBF with separate lengths
Geography	2	0.15 [0.02, 0.28]	0.41 [0.29, 0.51]
Socioeconomics	4	0.46 [0.35, 0.56]	0.47 [0.36, 0.56]
Socioeconomics + Geography	6	0.57 [0.48, 0.65]	0.60 [0.51, 0.68]
Demographics	7	0.41 [0.29, 0.51]	0.49 [0.39, 0.59]
Demographics + Geography	9	0.42 [0.31, 0.52]	0.56 [0.46, 0.64]
All non-language	13	0.61 [0.52, 0.68]	0.65 [0.57, 0.72]
Twitter, 10 PCA dimensions	10	0.28 [0.16, 0.40]	0.46 [0.35, 0.55]
+ All non-language	23	0.61 [0.52, 0.69]	0.66 [0.58, 0.73]
Twitter, 15 PCA dimensions	15	0.42 [0.31, 0.52]	0.53 [0.43, 0.62]
+ All non-language	28	0.64 [0.55, 0.71]	0.68 [0.60, 0.74]
Twitter, 25 PCA dimensions	25	0.51 [0.41, 0.60]	0.61 [0.52, 0.68]
+ All non-language	38	0.68 [0.61, 0.75]	0.68 [0.61, 0.75]
Twitter, 50 PCA dimensions	50	0.55 [0.45, 0.63]	0.62 [0.54, 0.70]
+ All non-language	63	0.67 [0.59, 0.74]	0.68 [0.60, 0.74]
Twitter, 100 PCA dimensions	100	0.56 [0.46, 0.64]	0.62 [0.53, 0.69]
+ All non-language	113	0.67 [0.59, 0.74]	0.67 [0.59, 0.74]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Appendix E.: Twitter interpolations

In Fig. E.3, we show the results of using reduced PCA dimensions from Twitter language, as opposed to Fig. 2(b) which combines Twitter and all non-language variables.

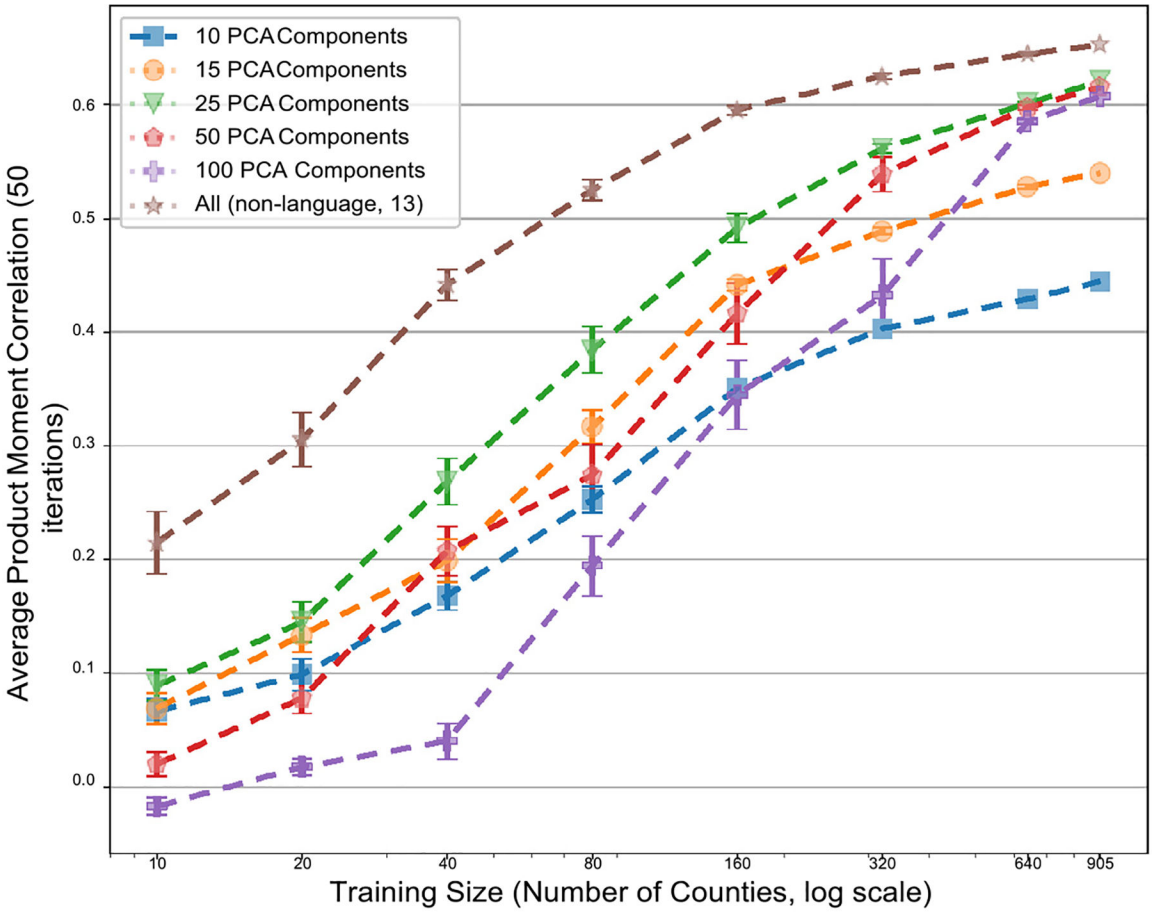

Fig. E.3.

Average out-of-sample prediction accuracy (product moment correlation) using only PCA reduced Twitter language. Unlike 2, where the models contain both language and non-language features, these models only contain Twitter features. Accuracies are average product moment correlation across 50 random training samples. Error bars are standard errors calculated across the 50 correlations.

References

Abebe R, Giorgi S, Tedijanto A, Buffone A, Schwartz HA, 2020. Quantifying community characteristics of maternal mortality using social media. In: The World Wide Web Conference.

Arora A, Spatz E, Herrin J, Riley C, Roy B, Kell K, Coberley C, Rula E, Krumholz HM, 2016. Population well-being measures help explain geographic disparities in life expectancy at the county level. Health Aff. 35 (11), 2075–2082.

Blei DM, Ng AY, Jordan MI, 2003. Latent Dirichlet allocation. J. Mach. Learn. Res 3, 993–1022.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Bleidorn W, Schönbrodt F, Gebauer JE, Rentfrow PJ, Potter J, Gosling SD, 2016. To live among like-minded others: exploring the links between person-city personality fit and self-esteem. Psychol. Sci 27 (3), 419–427. [PubMed: 26842317]

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. , 2020. Language Models Are Few-Shot Learners. Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901.

Chilès J-P, Desassis N, 2018. Fifty years of kriging. In: Handbook of Mathematical Geosciences: Fifty Years of IAMG, pp. 589–612.

Cressie N, 1990. The origins of kriging. Math. Geol 22 (3), 239–252.

Cui J, Zhang T, Jaidka K, Pang D, Sherman G, Jakhetiya V, Ungar LH, Guntuku SC, 2022. Social media reveals urban-rural differences in stress across China. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 114–124.

Culotta A, 2014. Estimating county health statistics with Twitter. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems. ACM, pp. 1335–1344.

Curtis B, Giorgi S, Buffone AE, Ungar LH, Ashford RD, Hemmons J, Summers D, Hamilton C, Schwartz HA, 2018. Can Twitter be used to predict county excessive alcohol consumption rates? PLoS ONE 13 (4), e0194290. [PubMed: 29617408]

Devlin J, Chang M-W, Lee K, Toutanova K, 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://aclanthology.org/N19-1423.

Diener E, Suh EM, Lucas RE, Smith HL, 1999. Subjective well-being: three decades of progress. Psychol. Bull 125 (2), 276.

Ebert T, Gebauer JE, Brenner T, Bleidorn W, Gosling S, Potter J, Rentfrow PJ, 2019. Are regional differences in personality and their correlates robust? Applying spatial analysis techniques to examine regional variation in personality across the US and Germany. Tech. Rep, Working papers on Innovation and Space.

Ebert T, Götz FM, Mewes L, Rentfrow PJ, 2023. Spatial analysis for psychologists: how to use individual-level data for research at the geographically aggregated level. Psychol. Methods 28 (5), 1100–1121. [PubMed: 35653725]

Edo-Osagie O, De La Iglesia B, Lake I, Edeghere O, 2020. A scoping review of the use of Twitter for public health research. Comput. Biol. Med 122, 103770. [PubMed: 32502758]

Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, Jha S, Agrawal M, Dziurzynski LA, Sap M, et al. , 2015. Psychological language on Twitter predicts county-level heart disease mortality. Psychol. Sci 0956797614557867.

Eichstaedt JC, Kern ML, Yaden DB, Schwartz H, Giorgi S, Park G, Hagan CA, Tobolsky VA, Smith LK, Buffone A, et al. , 2021. Closed- and open-vocabulary approaches to text analysis: a review, quantitative comparison, and recommendations. Psychol. Methods 26 (4), 398. [PubMed: 34726465]

Elleman LG, Condon DM, Holtzman NS, Allen VR, Revelle W, 2020. Smaller is better: associations between personality and demographics are improved by examining narrower traits and regions. Collabra, Psychol. 6 (1).

Gardner J, Pleiss G, Weinberger KQ, Bindel D, Wilson AG, 2018. Gpytorch: Black-box Matrix-Matrix Gaussian Process Inference with Gpu Acceleration. Advances in Neural Information Processing Systems, vol. 31, pp. 7576–7586.

Gibbons J, Malouf R, Spitzberg B, Martinez L, Appleyard B, Thompson C, Nara A, Tsou M-H, 2019. Twitter-based measures of neighborhood sentiment as predictors of residential population health. PLoS ONE 14 (7), e0219550. [PubMed: 31295294]

Giorgi S, Preotiuc-Pietro D, Buffone A, Rieman D, Ungar LH, Schwartz HA, 2018. The remarkable benefit of user-level aggregation for lexical-based population-level predictions. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Giorgi S, Lynn VE, Gupta K, Ahmed F, Matz S, Ungar LH, Schwartz HA, 2022a. Correcting sociodemographic selection biases for population prediction from social media. Proc. Int. AAAI Conf. Web Soc. Media 16 (1), 228–240. https://ojs.aaai.org/index.php/ICWSM/article/view/19287.

Giorgi S, Nguyen KL, Eichstaedt JC, Kern ML, Yaden DB, Kosinski M, Seligman ME, Ungar LH, Andrew Schwartz H, Park G, 2022b. Regional personality assessment through social media language. J. Pers 90 (3), 405–425. [PubMed: 34536229]

Grootendorst M, 2022. Bertopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint. arXiv :2203.05794.

Hartung J, Knapp G, Sinha BK, Sinha BK, 2008. Statistical Meta-Analysis with Applications, vol. 6. Wiley Online Library.

Hehman E, Calanchini J, Flake JK, Leitner JB, 2019. Establishing construct validity evidence for regional measures of explicit and implicit racial bias. J. Exp. Psychol. Gen 148 (6), 1022. [PubMed: 31180716]

Hoover J, Dehghani M, 2020. The big, the bad, and the ugly: geographic estimation with flawed psychological data. Psychol. Methods 25 (4), 412. [PubMed: 31647254]

Jaidka K, Giorgi S, Schwartz HA, Kern ML, Ungar LH, Eichstaedt JC, 2020. Estimating geographic subjective well-being from Twitter: a comparison of dictionary and data-driven language methods. Proc. Natl. Acad. Sci 117 (19), 10165–10171. 10.1073/pnas.1906364117. https://www.pnas.org/content/early/2020/04/24/1906364117. [PubMed: 32341156]

Kahneman D, Deaton A, 2010. High income improves evaluation of life but not emotional well-being. Proc. Natl. Acad. Sci 107 (38), 16489–16493. [PubMed: 20823223]

Landauer TK, Foltz PW, Laham D, 1998. An introduction to latent semantic analysis. Discourse Process. 25 (2–3), 259–284.

Lawless NM, Lucas RE, 2011. Predictors of regional well-being: a county level analysis. Soc. Indic. Res 101 (3), 341–357.

Lee H, Singh GK, 2020. Inequalities in life expectancy and all-cause mortality in the United States by levels of happiness and life satisfaction: a longitudinal study. Int. J. Matern. Child Health AIDS 9 (3), 305.

Matz SC, Gladstone JJ, 2020. Nice guys finish last: when and why agreeableness is associated with economic hardship. J. Pers. Soc. Psychol 118 (3), 545. [PubMed: 30307276]

Miranda Filho R, Almeida JM, Pappa GL, 2015. Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In: Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on. IEEE, pp. 1254–1261.

Moran PA, 1950. Notes on continuous stochastic phenomena. Biometrika 37 (1/2), 17–23. [PubMed: 15420245]

Ng YC, Colombo N, Silva R, 2018. Bayesian Semi-Supervised Learning with Graph Gaussian Processes. Advances in Neural Information Processing Systems, vol. 31.

Remington PL, Catlin BB, Gennuso KP, 2015. The county health rankings: rationale and methods. Popul. Health Metr 13 (1), 11. [PubMed: 25931988]

Rentfrow PJ, 2020. Geographical psychology. Curr. Opin. Psychol 32, 165–170. [PubMed: 31675524]

Rogers A, Kovaleva O, Rumshisky A, 2021. A primer in bertology: what we know about how BERT works. Trans. Assoc. Comput. Linguist 8, 842–866.

Schulz E, Speekenbrink M, Krause A, 2018. A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions. J. Math. Psychol 85, 1–16.

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Agrawal M, Park GJ, Lakshmikanth SK, Jha S, Seligman ME, Ungar L, 2013a. Characterizing geographic variation in well-being using tweets. In: ICWSM.

Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, Shah A, Kosinski M, Stillwell D, Seligman ME, et al. , 2013b. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS ONE 8 (9), e73791. [PubMed: 24086296]

Sibson R, 1981. A brief description of natural neighbor interpolation. In: Barnett V (Ed.), Interpreting Multivariate Data. John Wiley, Chichester, pp. 21–36 (Chapter 2).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Stelter M, Essien I, Sander C, Degner J, 2022. Racial bias in police traffic stops: white residents' county-level prejudice and stereotypes are related to disproportionate stopping of black drivers. Psychol. Sci 33 (4), 483–496. 10.1177/09567976211051272. [PubMed: 35319309]

Thomson R, Yuki M, Talhelm T, Schug J, Kito M, Ayanian AH, Becker JC, Becker M, Chiu C. y., Choi H-S, et al. , 2018. Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. Proc. Natl. Acad. Sci 115 (29), 7521–7526. [PubMed: 29959208]

Wadsworth T, Pendergast PM, 2014. Obesity (sometimes) matters: the importance of context in the relationship between obesity and life satisfaction. J. Health Soc. Behav 55 (2), 196–214. [PubMed: 24872467]

Ward G, De Neve J-E, Ungar LH, Eichstaedt JC, 2021. (Un)happiness and voting in US presidential elections. J. Pers. Soc. Psychol 120 (2), 370. [PubMed: 32700960]

Williams CK, Rasmussen CE, 2006. Gaussian Processes for Machine Learning, vol. 2. MIT Press, Cambridge, MA.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

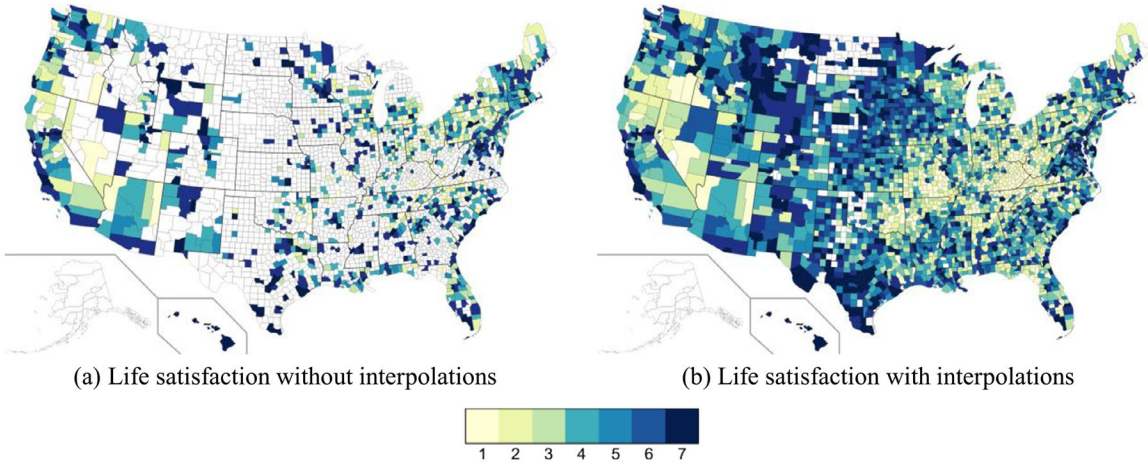

Fig. 1.

U.S. county map of Gallup life satisfaction (seven quantiles) using (a) 1,133 counties with at least 300 individual-level responses and (b) 2,954 counties that include both the counties that meet the minimum response threshold plus counties with interpolated life satisfaction. Higher quantile number indicates greater life satisfaction. White cells contain no data as several counties did not have data available for all measures (features) and, thus, interpolation was not possible.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

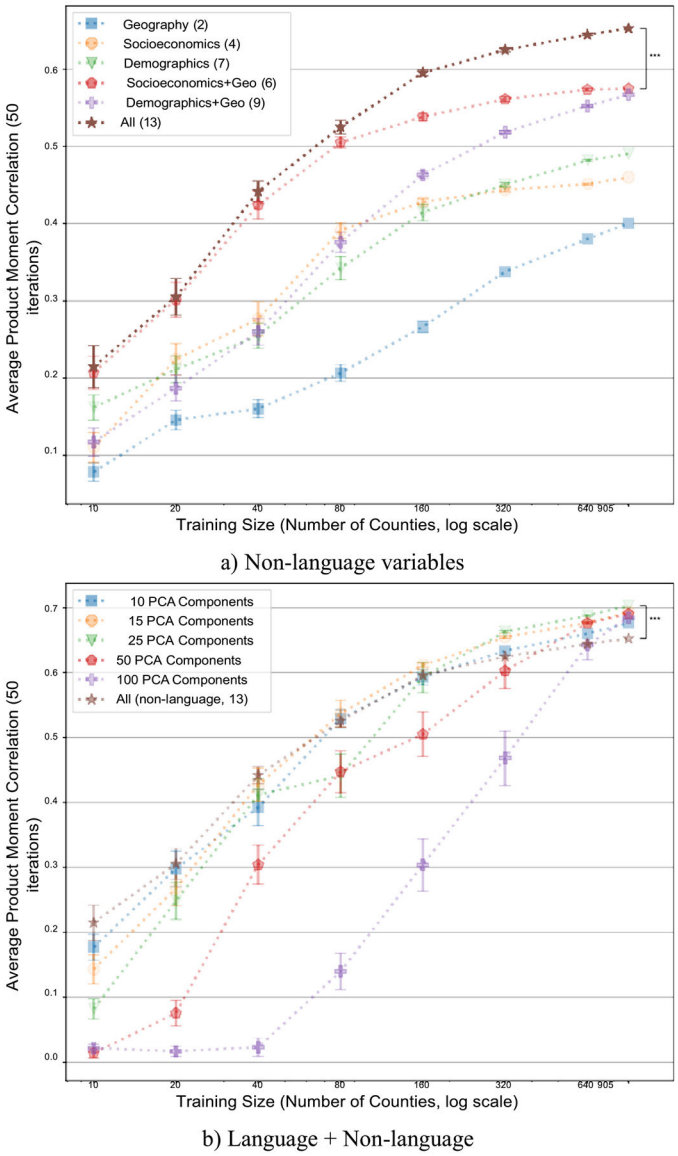

Fig. 2.

Life satisfaction prediction accuracies as a function of varying training sample sizes (x-axis) and feature spaces (lines): (a) non-language variables and (b) language combined with non-language variables. Accuracies are average product moment correlation across 50 random training samples. Error bars are standard errors calculated across the 50 correlations. *** significant difference (paired t-test; $p < 0.001$) between model accuracies: (a) all non-language features vs. socioeconomics + geography and (b) 25 PCA Twitter components + all non-language features (the top performing Twitter model) vs. all non-language features.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Out-of-sample prediction accuracy (product moment correlations with 95% confidence intervals). All correlations significant at $p < 0.001$.

	Number of Features	Test Set Correlation
Geography	2	0.40 [0.29, 0.50]
Socioeconomics	4	0.46 [0.35, 0.56]
Socioeconomics + Geography	6	0.57 [0.48, 0.66]
Demographics	7	0.49 [0.39, 0.58]
Demographics + Geography	9	0.57 [0.47, 0.65]
All non-language	13	0.65 [0.57, 0.72]
Twitter, 10 PCA components	10	0.44 [0.33, 0.54]
+ All non-language	23	0.68 [0.60, 0.74]
Twitter, 15 PCA components	15	0.54 [0.44, 0.63]
+ All non-language	28	0.69 [0.62, 0.76]
Twitter, 25 PCA components	25	0.62 [0.53, 0.69]
+ All non-language	38	0.70 [0.63, 0.76]
Twitter, 50 PCA components	50	0.62 [0.53, 0.69]
+ All non-language	63	0.69 [0.62, 0.75]
Twitter, 100 PCA components	100	0.61 [0.52, 0.68]
+ All non-language	113	0.68 [0.61, 0.75]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Combining GP interpolations with noisy data. Reported mean, standard error, and 95% confidence intervals of the product moment correlation across 50 iterations. Standard error is calculated on the 50 correlations (to measure the variation in effect size due to the random sampling of participant-level data), while the 95% confidence intervals are calculated for the mean correlation. We use the best performing model from Fig. 2(b): 25 Twitter PCA components + all non-language features. Models trained on the training data set (905 counties) and evaluated on the test data set (228 counties). Average Life Sat. is the average life satisfaction estimate across a random sample with corresponding sample size.

	Participant Level Sample Size			
	25	**50**	**100**	**200**
GP Interpolation	0.70 (0.000) [0.69, 0.72]	0.70 (0.000) [0.69, 0.72]	0.70 (0.000) [0.69, 0.72]	0.70 (0.000) [0.69, 0.72]
Average Life Sat.	0.38 (0.007) [0.35, 0.41]	0.49 (0.007) [0.47, 0.52]	0.65 (0.005) [0.63, 0.67]	0.79 (0.003) [0.78, 0.80]
Combined	0.73 (0.002) [0.72, 0.75]	0.75 (0.002) [0.74, 0.77]	0.80 (0.002) [0.78, 0.81]	0.85 (0.002) [0.84, 0.86]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Validation with external criteria: product moment correlations between gold standard life satisfaction and external variables across high-count counties alone (i.e., non-interpolated life satisfaction from all counties in the train and test data sets; a baseline measure of the relationship between life satisfaction and external criteria) and all counties (i.e., a combination of high- and low-count counties; to see how our methods change the relationship between life satisfaction and external criteria). We use the best performing model from Fig. 2(a): all non-language features.

	N	Life Expectancy	Obesity	Income[†]	Education[‡]
High-count counties					
Average Life Sat.	1133	0.52 [0.47, 0.56]	−0.43 [−0.48, −0.39]	0.42 [0.37, 0.47]	0.38 [0.33, 0.43]
All counties					
State average	2954	0.32 [0.29, 0.35]	−0.29 [−0.32, −0.25]	0.28 [0.25, 0.32]	0.21 [0.18, 0.25]
GP Interpolation	2954	0.50 [0.47, 0.52]	−0.39 [−0.42, −0.36]	0.37 [0.34, 0.41]	0.30 [0.27, 0.33]
Average Life Sat.	2954	0.35 [0.32, 0.38]	−0.26 [−0.29, −0.23]	0.31 [0.28, 0.35]	0.23 [0.20, 0.27]
Combined	2954	0.51 [∇] [0.48, 0.53]	−0.39 [∇] [−0.42, −0.36]	0.40 [∇] [0.37, 0.43]	0.31 [∇] [0.28, 0.34]

[∇] bootstrap test shows *no* significant difference in effect size when compared to average life satisfaction across the high-count counties,

[†] income removed from the interpolation feature space,

[‡] education removed from the interpolation feature space.