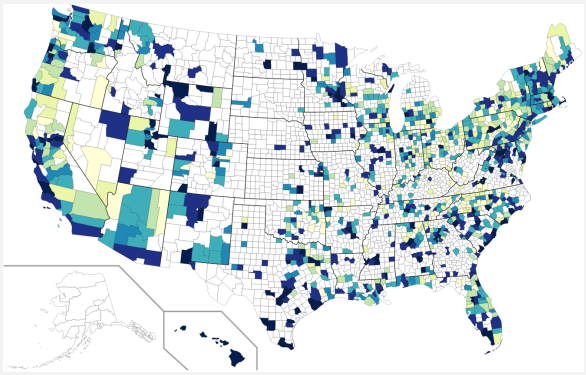
Supplementary Material to:

Towards Well-Being Measurement with Social Media Across Space, Time and Cultures: Three Generations of Progress

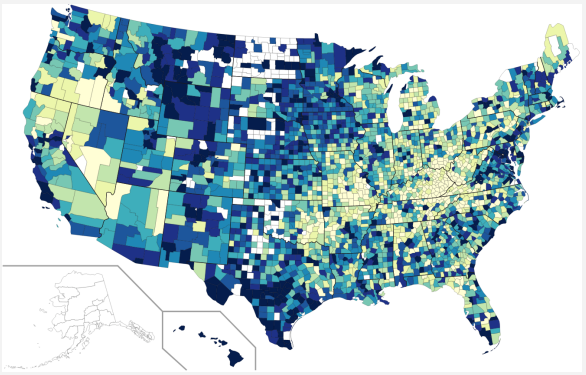
Spatial Interpolation

Recent advances in machine learning and geo-statistics have paved the way for more sophisticated spatial interpolation techniques, which allow for near-full national coverage of psychological constructs at a fine-grained resolution. For example, traditional implementations of Gaussian Processes (i.e., standard models used for spatial interpolation) rely on manually-chosen model hyperparameters, whereas modern implementations *learn* these hyperparameters from the data. This is especially important when using high-dimensional data, which increases the number of hyperparameters in the model. These methods can leverage both high dimensional, non-proximity-based community characteristics such as sociodemographic markers or Twitter language data (i.e., measures of socioeconomic and cultural similarity) alongside standard spatial proximity (e.g., neighboring counties) to interpolate psychological constructs where data sparsity prohibits community measurements (Giorgi et al., under review). This allows researchers to measure psychological constructs at scale and gain insights into communities that may have been inaccessible via traditional survey-based measures, for example, sparsely populated rural areas. In an evaluation, we found that using this interpolation yielded life satisfaction estimates for 3,000 US counties that showed the same correlations with external criteria (e.g., health and socioeconomics) as a smaller sample of 1,100 counties using non-interpolated survey-based life satisfaction (see figures below).

These methods are a powerful complement to a social-media-based estimation of small areas (such as counties), as they leverage existing data sources (demographics, socioeconomics) and principled measures of uncertainty to "fill in the map" for parts of countries in which populations, and thus social media posting, is more sparse (see also **Fig 1C**, where these methods were applied).



(A) Gallup life satisfaction



(**B**) Gallup life satisfaction interpolated using Gaussian Processes.

Spatial resolution below the county level

Among Twitter users, a recent estimate suggests that 12.9% of Tweets can be associated with GPS coordinates (and a total of 24.4% with geographic location, more generally; Huang & Carley, 2019). Therefore, aggregation to – in principle, arbitrarily small– geographic units (such as blocks or Census tracts) is possible. Cao et al., (2018) use *Gen 1* aggregation techniques combined with *Level 2* sentiment analysis (using the IBM Watson Alchemy API) to study spatiotemporal variations of individuals' tweets within Massachusetts. Language use was analyzed across land-use areas (e.g., forest, residential, commercial, golf course, marina) and time (including weekdays and time of day). Results revealed that individuals post tweets with more positive sentiment in commercial and public areas, around noon/evening, and during weekends. On the other hand, negative sentiments were more likely within farmland, transportation, and industrial areas, at midnight and during weekdays. Hence, the high resolution of big, ecologically valid data from social media enables answering detailed research questions about human activity and well-being.

Further examples of international Gen 1 Level 1 studies

Gen 1 was used in an exploratory study in Europe to examine nine European countries, including Germany, France, Sweden, Portugal, the Netherlands, Italy, Spain, Turkey, and the U.K., in their national languages over six years (Coşkun & Ozturan, 2018). Two hundred fifty-five million tweets from more than 110,000 users were analyzed (with sparse methodological detail reported). In general, using dictionary-based (*Level 1*) approaches applied to random Twitter samples (*Gen 1*) has been the most common across labs and research groups, but in terms of validation beyond an inspection of time series has only had mixed success in the literature.

The combination of post-stratification with sociodemographic estimates from social media has also been used in multilingual and multi-country settings (Wang et al., 2019). Wang et al. estimate age and gender using a deep neural network trained on profile images and text from 26 European countries and 32 languages to demographically characterize a multilingual Twitter sample to be used for estimation. This was achieved using a sample of 3 million Twitter users geolocated to NUTS3 regions (Nomenclature of Territorial Units for Statistics - 3), which is the finest-grained spatial resolution across E.U. member states across 26 countries. Age and gender estimates are then used with post-stratification methods to remove selection biases in the sample. The result is a large sample of Twitter users representative of heterogeneous regions across most of Europe, which can further be used in downstream applications. Together, these results show that samples from social media data can be aggregated in ways that accurately represent diverse, multilingual populations across small geographic areas.

Geographic and temporal predictions pose different difficulties

Language differ across regions. This includes examples like *soda* (used in the northeast U.S.) and *pop* (used in the Midwest U.S.) or eagles which can refer to a professional sports team in Philadelphia. In an early example of *Gen 2* person-centered methods, Eisenstein et al. (2010) proposed a data-driven topic model which jointly estimates both topics (semantically related clusters of words) and their regional geographic

variation, noting that words and regions interact to drive lexical frequencies. For example, standard topic modeling techniques may discover a high-level "sports" topic, whereas the methods proposed by Eisenstein et al. estimate "sports" topics differently across regions: topics that included *Boston* and *Celtics* were found in the northeast U.S., while topics that included *Kobe* and *Lakers* (i.e., players and teams in Los Angeles) were found in the southwest U.S. These methods also identify geographically-coherent lexical regions that varied both within and across standard regional borders, for example, inter- and intra-state linguistic variation. Within the state of California, the authors were able to distinguish northern California from southern California across several topics such as emoticon use and internet slang (e.g., local colloquialisms like *hella* in the north and *messin* in the south). This approach allows one to identify regional lexical variation beyond standard physical distance and nation-state borders, resulting in culturally local text representations. Furthermore, the data source itself (naturally occurring, human-generated social media text) is a unique ecological data source for identifying regional or cultural lexical variation, as compared to more formal types of text.

References

Cao, X., MacNaughton, P., Deng, Z., Yin, J., Zhang, X., & Allen, J. G. (2018). Using twitter to better understand the spatiotemporal patterns of public sentiment: A case study in Massachusetts, USA. *International Journal of Environmental Research and Public Health*, *15*(2), 250.

Coşkun, M., & Ozturan, M. (2018). # europehappinessmap: A framework for multi-lingual sentiment analysis via social media big data (a Twitter case study). *Information*, *9*(5), 102.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. (2010). A latent variable model for geographic lexical variation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277–1287.

Huang, B., & Carley, K. M. (2019). A large-scale empirical study of geotagging behavior on twitter. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 365–373.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic inference and representative population estimates from multilingual social media data. *The World Wide Web Conference*, 2056–2067.