The Journal of Open Source Software

DeGAUSS: Decentralized Geomarker Assessment for Multi-Site Studies

Cole Brokamp[1, 2]

1 Cincinnati Children's Hospital Medical Center **2** University of Cincinnati

DOI: 10.21105/joss.00812

Software

- Review
- Repository
- Archive

Submitted: 29 June 2018
Published: 02 October 2018

Summary

Scientific studies often examine the relationships between place-based information and health outcomes; for example, air pollution and asthma, neighborhood crime and mental health, or community greenspace and IQ. Study subjects with location information, most commonly a residential mailing address, are linked to databases of place-based information, or "geomarkers" in order to conduct these studies. Defined formally, a "geomarker" is any objective, contextual or geographic measure that influences or predicts the incidence of outcome or disease. "Geocoding" is the process of translating a string of text referring to a location (most often a mailing address) into coordinates on the earth's surface (most often latitude and longitude). These coordinates are required to link participants to their estimated exposures to geomarkers – a process we call "geomarker assessment". Some examples of geomarker assessment commonly performed in health studies using people include distance to the nearest major roaday – a commonly used as a measure of estimated exposure to traffic related air pollution that is associated with increased risk of asthma – or neighborhood median household income – a commonly used as a measure of community deprivation associated with increased bed days spent in the hospital.

Studies that utilize geocoding and assessment of any geomarkers frequently utilize residential addresses or other geolocation data that are considered protected health information (PHI). The HIPAA privacy rule (United States Public Law, 1996), the HITECH Act of 2009 (United States Public Law, 2009), and the Federal Policy for the Protection of Human Subjects (United States Public Law, 1981) establish regulations to safeguard the confidentiality of patients and research subjects when health care providers or researchers use PHI. While beneficial with repsect to privacy, this presents an outstanding challenge for researchers by preventing them from using external third party software to analyze and extract information from study participants' adddresses or locations. Furthermore, this restricts scientists' ability to collaborate by combining datasets containing any PHI. We are critically missing standard ways to make this easy.

`DeGAUSS` is a standalone, container-based application that can produce geocodes and derive community and environmental exposures. Usable on PC, Mac, or Linux machines, identifying information never leaves the local machine. Figure 1 illustrates the process of using `DeGAUSS` within a multi-site study. Each study site uses `DeGAUSS` to both independently geocode their own addresses and link in the necessary place-based characteristics. After any PHI is removed, the data including the geomarkers are no longer considered PHI and are available for sharing and collaboration. In addition to securing PHI, this guarantees that the software will always run the same, regardless of its environment, which is a vital requirement for reproducible research.

`DeGAUSS` relies heavily on `R` (R Core Team, 2014) and the geospatial packages `sp` (RS Bivand, Pebesma, & Gomez-Rubio, 2005), `rgdal` (Roger Bivand, Keitt, & Rowlingson,

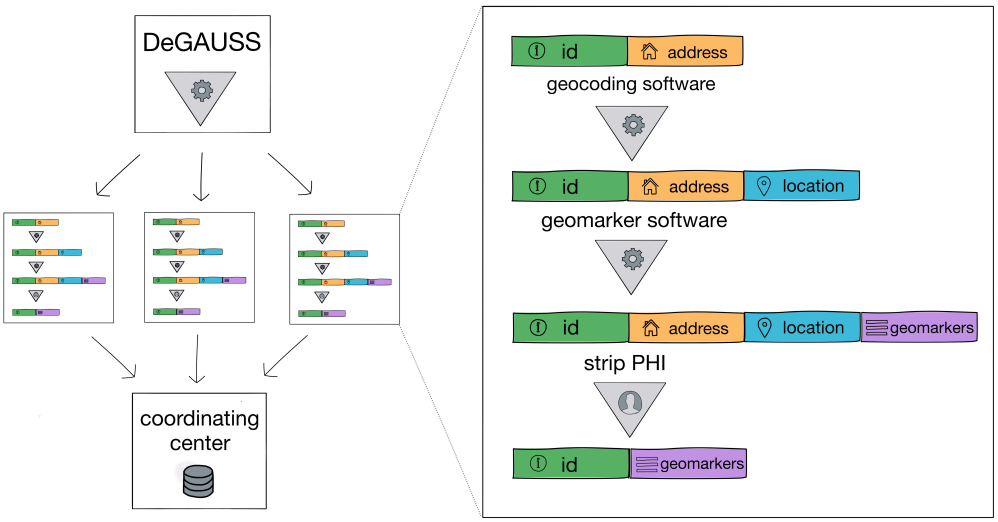

Figure 1: Illustration of DeGAUSS application within a multi-site study. Each site individually uses PHI to geocode and assign geomarkers. After PHI is removed, the dataset can be shared or combined with other datasets for analysis.

2014) `tigris` (Walker, 2017), and `tidycensus` (Walker, 2018). The underlying geocoder is based on the `usaddress` geocoder (Brokamp, 2017a). It was designed to be used by scientific researchers who wish to collect place-based data on study subjects and patients with a residential address. A proof of concept of the application of `DeGAUSS` within a multi-site study has previously been described (Brokamp, Wolfe, Lingren, Harley, & Ryan, 2018) and this approach is currently being adopted by several other multi-site cohort studies. Additionally, `DeGAUSS` has found use within the electronic health records of healthcare systems to automate geocoding and assessment of community characteristics and environmental exposures.

`DeGAUSS` is currently licensed under GNU GPLv3, archived on Zenodo with a linked DOI (Brokamp, 2017b), and is maintained on GitHub (https://github.com/cole-brokamp/DeGAUSS) where users can submit issues and propose their own extensions and additions.

References

Bivand, R., Keitt, T., & Rowlingson, B. (2014). *Rgdal: Bindings for the geospatial data abstraction library*. Retrieved from http://CRAN.R-project.org/package=rgdal

Bivand, R., Pebesma, E., & Gomez-Rubio, V. (2005). Classes and methods for spatial data in r. *R News*, *5*(9).

Brokamp, C. (2017a, March). geocoder: v2.2. doi:10.5281/zenodo.344621

Brokamp, C. (2017b, May). Cole-brokamp/degauss v0.2. doi:10.5281/zenodo.570873

Brokamp, C., Wolfe, C., Lingren, T., Harley, J., & Ryan, P. (2018). Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies. *Journal of the American Medical Informatics Association*, *25*(3), 309–314. doi:10.1093/jamia/ocx128

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

United States Public Law. (1981). Federal Policy for the Protection of Human Subjects ("Common Rule"). 45 CFR part 46.

United States Public Law. (1996). Health Insurance Portability and Accountability Act of 1996 (HIPAA) Pub.L. 104–191 and the HIPAA Privacy Rule 2003. 45 CFR Part 160 and Part 16 Subparts A and E.

United States Public Law. (2009). Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009. Section 13410(d).

Walker, K. (2017). *Tigris: Load census tiger/line shapefiles into r*. Retrieved from https://CRAN.R-project.org/package=tigris

Walker, K. (2018). *Tidycensus: Load us census boundary and attribute data as 'tidyverse' and 'sf'-ready data frames*. Retrieved from https://CRAN.R-project.org/package=tidycensus