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.00800

Software

- Review
- Repository
- Archive

Submitted: 29 June 2018
Published: 02 July 2018

Summary

Health studies that utilize geocoding and assessment of any place-based characteristic, e.g. census-tract level income or estimated exposure to ambient air pollution, 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. This is an outstanding challenge for multi-site studies because current approaches include getting institution-specific approval for a central site to conduct all analyses, which is a lengthy and sometime unfeasible process, or allowing each site to conduct their own analyses which requires expertise at each site and can result in non-reproducible and inconsistent geocoding and assessment of place-based characteristics, i.e. geomarkers.

`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, strips any PHI, and then sends de-identified dataset out for analysis. 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` (Bivand, Pebesma, and Gomez-Rubio 2005), `rgdal` (Bivand, Keitt, and Rowlingson 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 et al. 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.

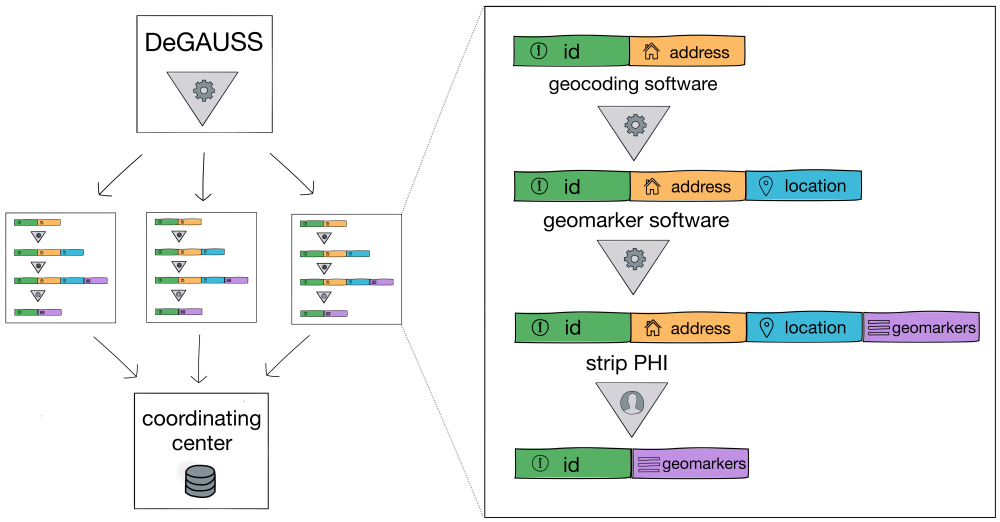

Figure 1: Illustration of DeGAUSS application within a multi-site study.

References

Bivand, Roger, Tim Keitt, and Barry Rowlingson. 2014. *Rgdal: Bindings for the Geospatial Data Abstraction Library*. http://CRAN.R-project.org/package=rgdal.

Bivand, RS, EJ Pebesma, and V Gomez-Rubio. 2005. "Classes and Methods for Spatial Data in R." *R News* 5 (9).

Brokamp, Cole. 2017a. "geocoder: v2.2." https://doi.org/10.5281/zenodo.344621.

———. 2017b. "Cole-Brokamp/Degauss V0.2." https://doi.org/10.5281/zenodo.570873.

Brokamp, Cole, Chris Wolfe, Todd Lingren, John Harley, and Patrick Ryan. 2018. "Decentralized and Reproducible Geocoding and Characterization of Community and Environmental Exposures for Multisite Studies." *Journal of the American Medical Informatics Association* 25 (3). Oxford University Press (OUP):309–14. https://doi.org/10.1093/jamia/ocx128.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

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

———. 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."

———. 2009. "Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009. Section 13410(d)."

Walker, Kyle. 2017. *Tigris: Load Census Tiger/Line Shapefiles into R*. https://CRAN.R-project.org/package=tigris.

———. 2018. *Tidycensus: Load Us Census Boundary and Attribute Data as 'Tidyverse' and 'Sf'-Ready Data Frames*. https://CRAN.R-project.org/package=tidycensus.