Areal Interpolation Methods

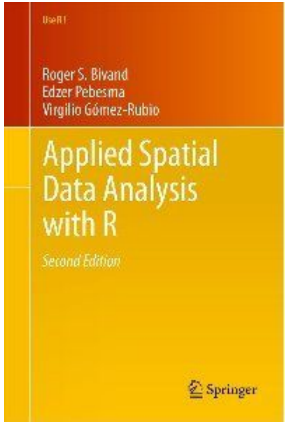
Edzer Pebesma
Institute for Geoinformatics, University of Münster
UseR! 2026, Jul 6-9 2026, Warsaw, PL

Slides: https://github.com/edzer/UseR2026

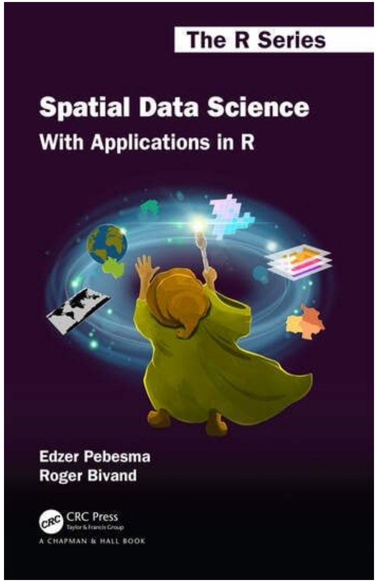
ifgi

Spatial Data Science

Spatial Data Science (or: GI Science, geocomputation, geospatial data science, geographic data science, spatial computing, etc.)



2008, 2013



2023

Spatial Data Science

With Applications in R

AUTHORS

Edzer Pebesma

Roger Bivand

AFFILIATIONS

Universität Münster

Norwegian School of Economics

PUBLISHED

January 22, 2025

Preface

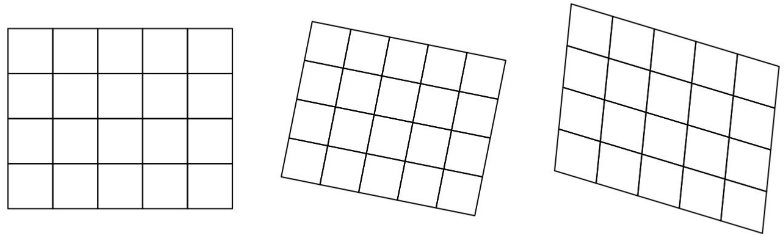
Data science is concerned with finding answers to questions on the basis of available data, and communicating that effort. Besides showing the results, this communication involves sharing the data used, but also exposing the path that led to the answers in a comprehensive and reproducible way. It also acknowledges the fact that available data may not be sufficient to answer questions, and that any answers are conditional on the data collection or sampling protocols employed.

Spatial data

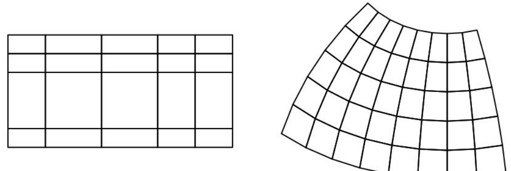
Formats:

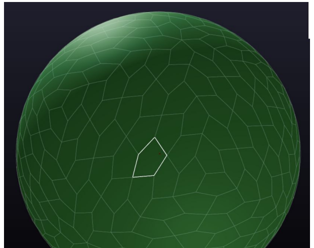
- Vector geoms: points/lines/polygons
- Raster
- Arrays

regular rotated sheared



rectilinear curvilinear





Phenomena:

- Spatially continuous (fields, coverages)
- Spatially discrete (objects, events)

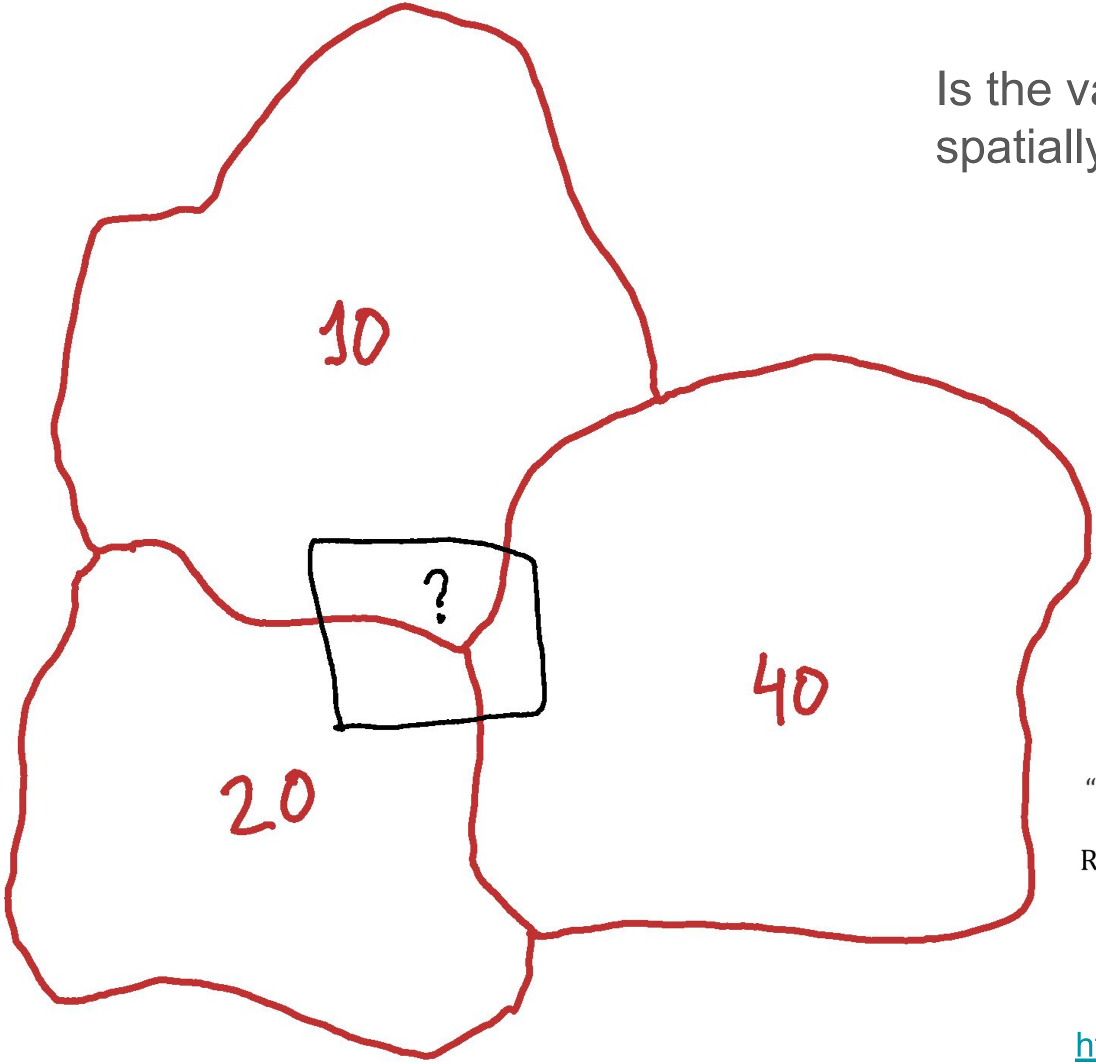
Data generation:

- Sampling
- Exhaustive observation over an observation window (window = sample)
- Aggregation over an area (count, sum, mean, max); area may be discrete or kernel

Support, attribute-geometry relationship

After (re)projection, redistricting (spatial aggregation or disaggregation, up- or downscaling) is the second largest SDS challenge (source of error) in spatial data science.

- Q1: does a variable associated with a *set* of points (lines, polygon, multipoint or raster cell geometries) refer to each point or to the set?
    - Each point: land cover, lithology, soil type -> point support
    - The set: population, population density, average temperature -> “block” support
- Q2: is the “block” support variable spatially extensive or intensive?
    - What happens with the variable if we split a geometry in two?
- If our data sources do not contain this information, how can we avoid user error?



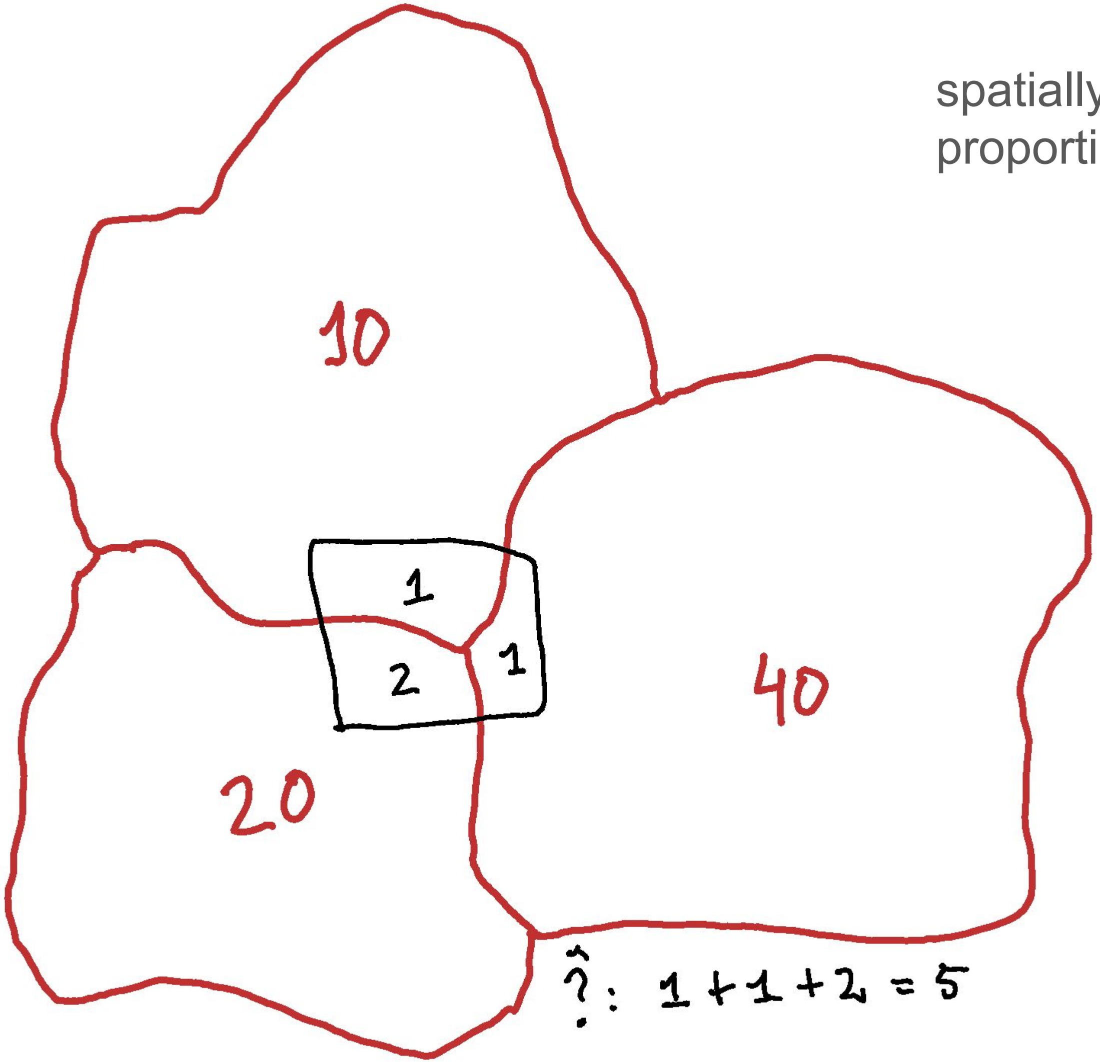
Is the variable spatially *extensive* or spatially *intensive*?

“Spatial reallocation of areal data – another look at basic methods
Réaffectation spatiale de données surfaciques - une autre approche des méthodes de base”

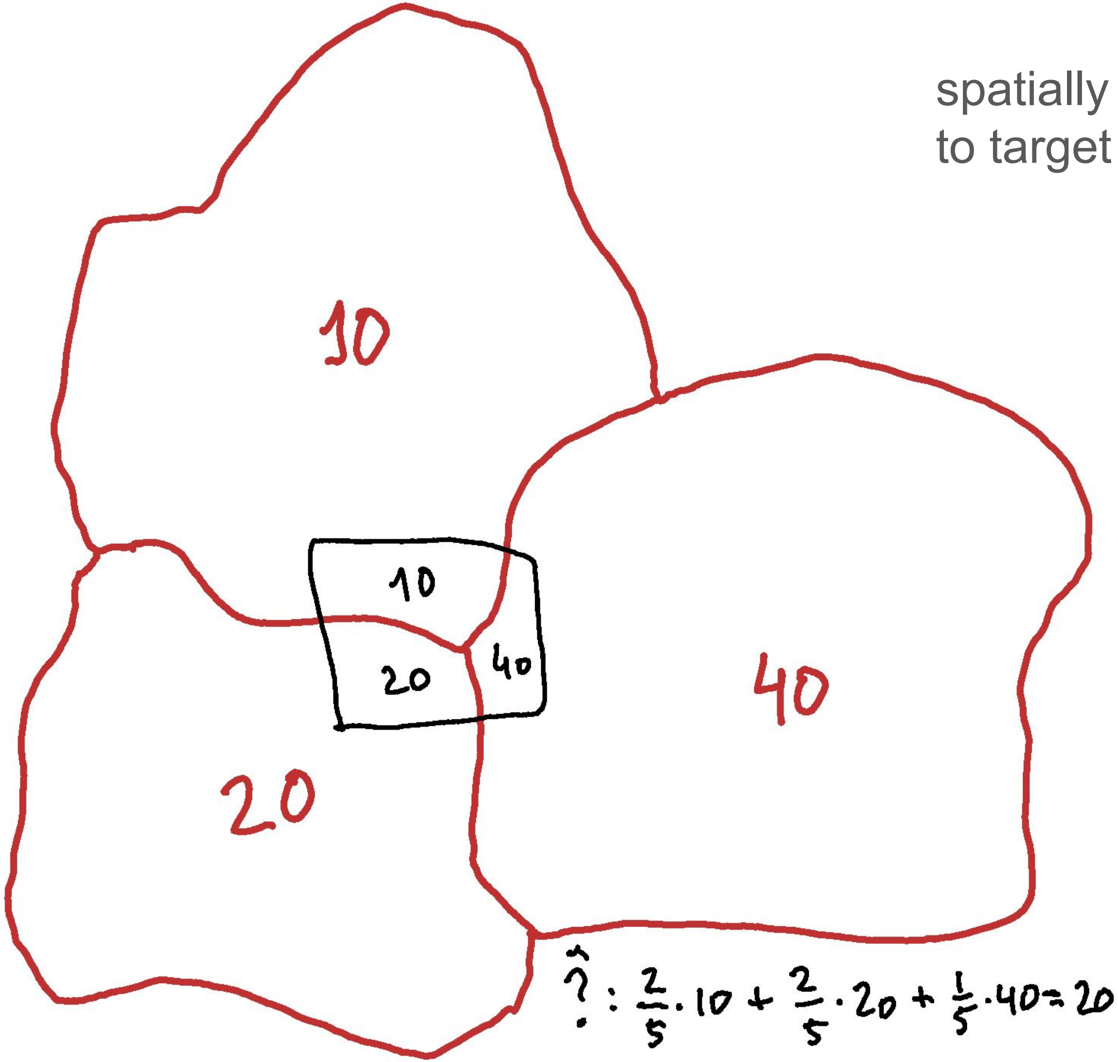
V. H. Do, C. Thomas-Agnan, A. Vanhems

http://www.thibault.laurent.free.fr/code/areal/

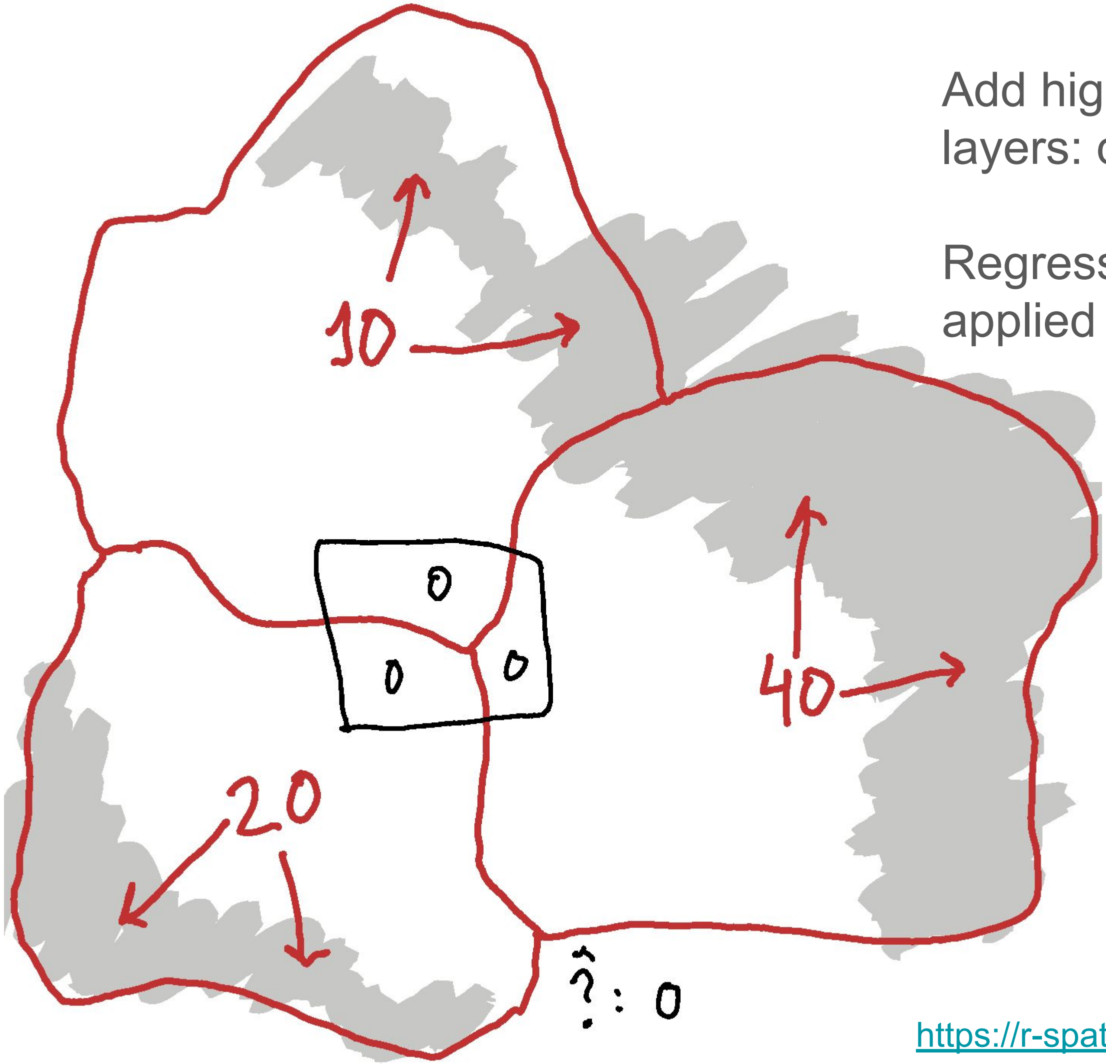
spatially *extensive*: redistribute proportional to source areas (red)



spatially *intensive*: weigh, proportional to target area (black)



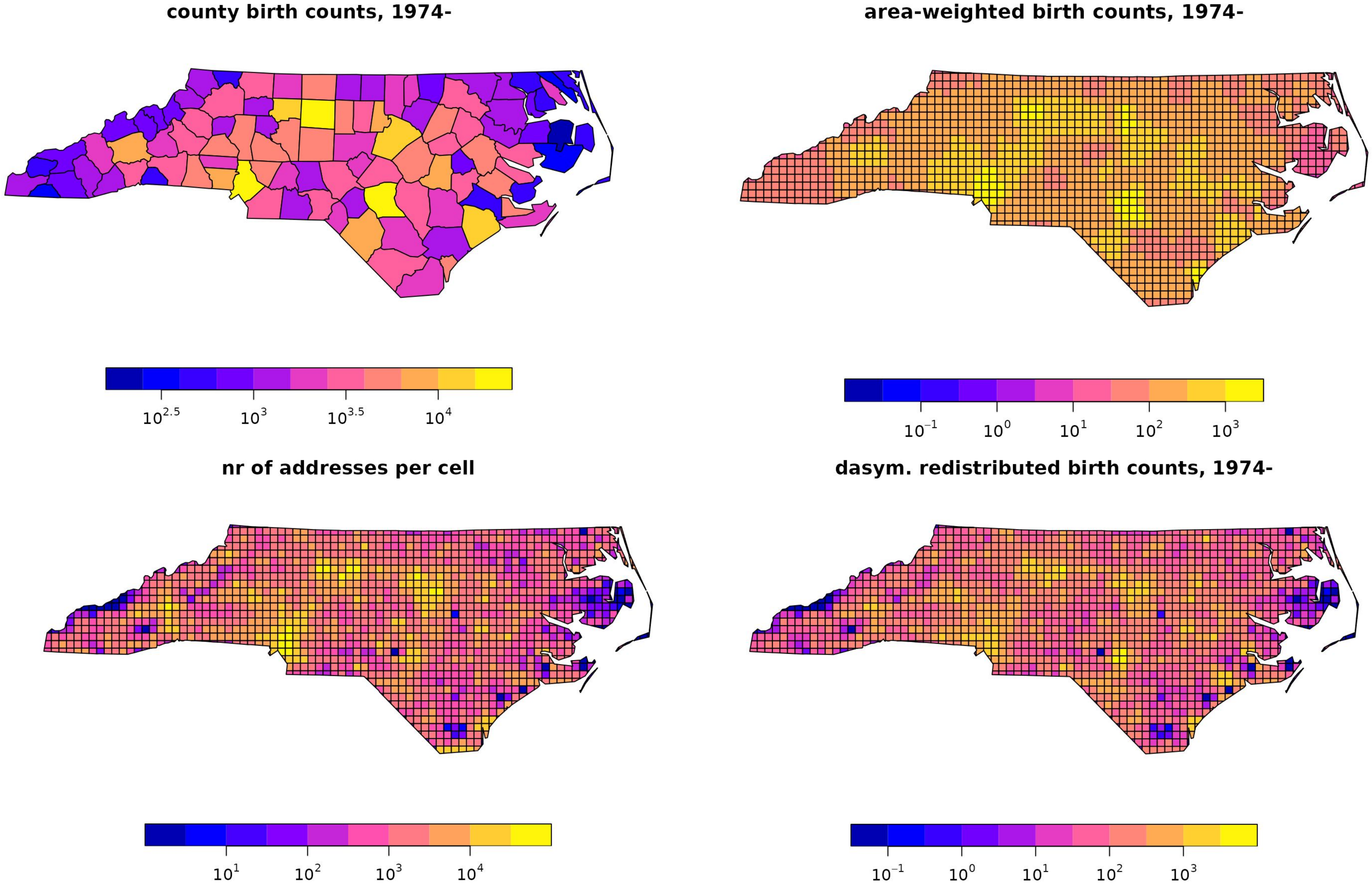
$$\hat{?}: \frac{2}{5}\cdot 10 + \frac{2}{5}\cdot 20 + \frac{1}{5}\cdot 40 = 20$$



Add high(er) resolution, informative layers: dasymetric mapping

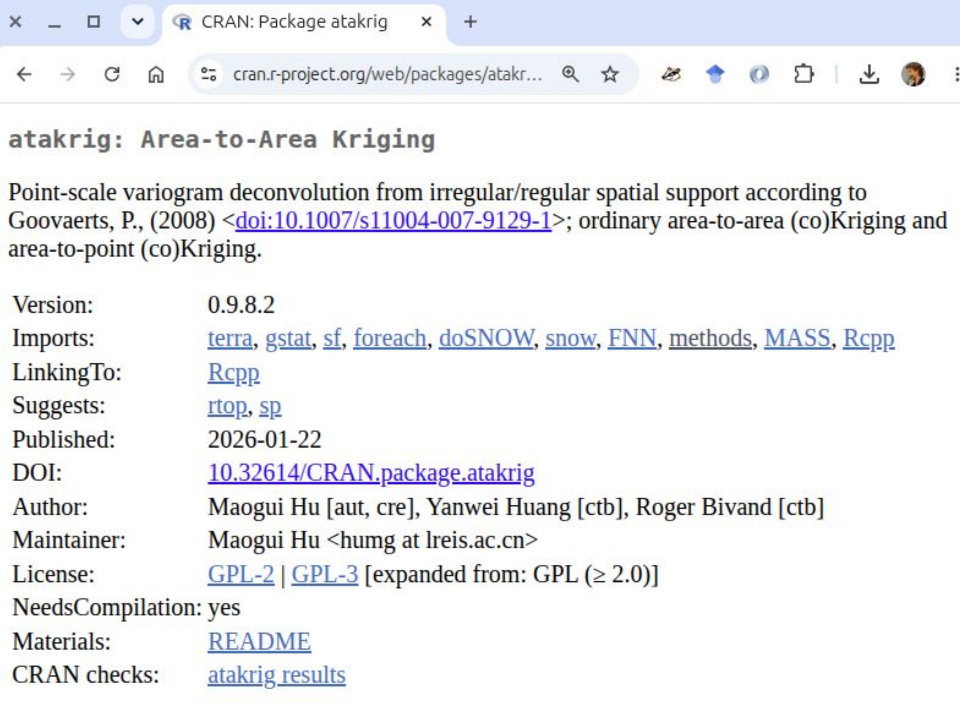
Regression methods have been applied to smooth noise

https://r-spatial.github.io/sf/reference/interpolate_aw.html



county birth counts, 1974-
$10^{2.5}$
10^{3}
$10^{3.5}$
10^{4}
area-weighted birth counts, 1974-
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
nr of addresses per cell
10^{1}
10^{2}
10^{3}
10^{4}
dasym. redistributed birth counts, 1974-
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}

Geostatistical approaches: “use all data”



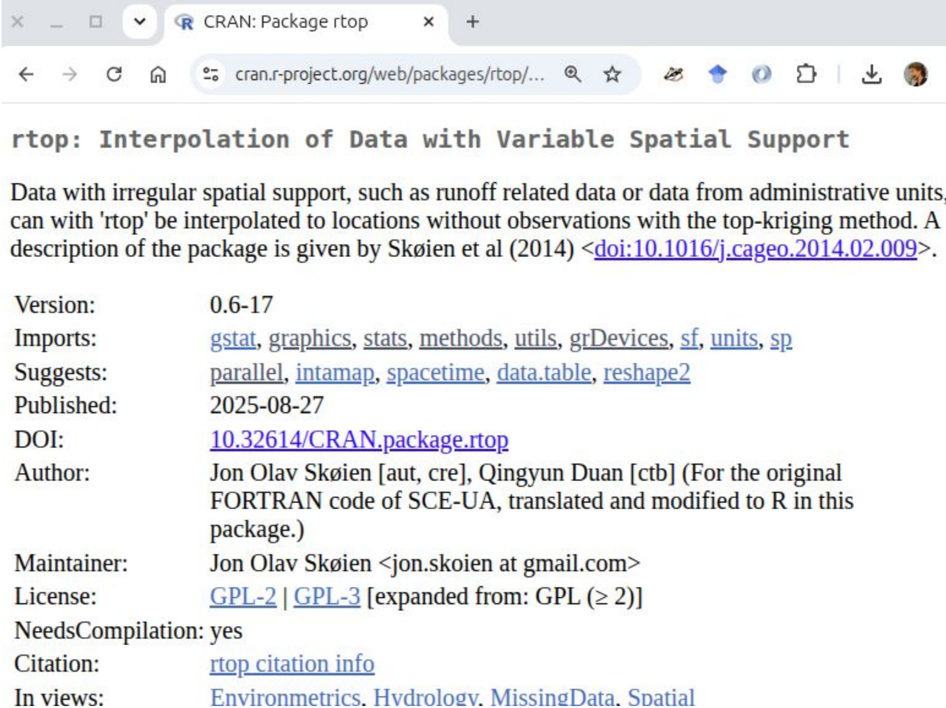
CRAN: Package atakrig

cran.r-project.org/web/packages/atakr...

atakrig: Area-to-Area Kriging

Point-scale variogram deconvolution from irregular/regular spatial support according to Goovaerts, P., (2008) <doi:10.1007/s11004-007-9129-1>; ordinary area-to-area (co)Kriging and area-to-point (co)Kriging.

Version:	0.9.8.2
Imports:	terra, gstat, sf, foreach, doSNOW, snow, FNN, methods, MASS, Rcpp
LinkingTo:	Rcpp
Suggests:	rtop, sp
Published:	2026-01-22
DOI:	10.32614/CRAN.package.atakrig
Author:	Maogui Hu [aut, cre], Yanwei Huang [ctb], Roger Bivand [ctb]
Maintainer:	Maogui Hu <humg at lreis.ac.cn>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2.0)]
NeedsCompilation:	yes
Materials:	README
CRAN checks:	atakrig results



CRAN: Package rtop

cran.r-project.org/web/packages/rtop/...

rtop: Interpolation of Data with Variable Spatial Support

Data with irregular spatial support, such as runoff related data or data from administrative units, can with 'rtop' be interpolated to locations without observations with the top-kriging method. A description of the package is given by Skøien et al (2014) <doi:10.1016/j.cageo.2014.02.009>.

Version:	0.6-17
Imports:	gstat, graphics, stats, methods, utils, grDevices, sf, units, sp
Suggests:	parallel, intamap, spacetime, data.table, reshape2
Published:	2025-08-27
DOI:	10.32614/CRAN.package.rtop
Author:	Jon Olav Skøien [aut, cre], Qingyun Duan [ctb] (For the original FORTRAN code of SCE-UA, translated and modified to R in this package.)
Maintainer:	Jon Olav Skøien <jon.skoien at gmail.com>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
NeedsCompilation:	yes
Citation:	rtop citation info
In views:	Environmetrics, Hydrology, MissingData, Spatial

- Requires “point-point” covariance (“deconvolution”)
- Requires area-area mean covariance, which involves computing a 4D integral

Statistical approaches

Journal of Statistical Software

April 2023, Volume 106, Issue 11. doi: 10.18637/jss.v106.i11

disaggregation: An R Package for Bayesian Spatial Disaggregation Modeling

Anita K. Nandi
University of Oxford

Tim C. D. Lucas
University of Oxford

Rohan Arambepola
University of Oxford

Peter Gething
Telethon Kids Institute
Curtin University

Daniel J. Weiss
University of Oxford

Abstract

Disaggregation modeling, or downscaling, has become an important discipline in epidemiology. Surveillance data, aggregated over large regions, is becoming more common, leading to an increasing demand for modeling frameworks that can deal with this data to understand spatial patterns. Disaggregation regression models use response data aggregated over large heterogeneous regions to make predictions at fine-scale over the region by using fine-scale covariates to inform the heterogeneity. This paper presents the R package **disaggregation**, which provides functionality to streamline the process of running a disaggregation model for fine-scale predictions.

Keywords: Bayesian spatial modeling, disaggregation modeling, TMB.

$$\mathrm{link}(\mathrm{rate}_{ij}) = \beta_0 + \beta X_{ij} + \mathrm{GP}(s_{ij}) + u_i$$

$$\mathrm{cases}_i = \sum_{j=1}^{N_i} a_{ij}\mathrm{rate}_{ij}$$

$$\mathrm{rate}_i = \frac{\mathrm{cases}_i}{\sum_{j=1}^{N_i} a_{ij}}$$

$$y_i \mid \beta_0, \beta_i, \mathrm{GP}, u_i \sim \mathrm{Poisson}(\mathrm{cases}_i)$$

- Uses INLA for linear link
- TMB for L.A. on other links
- GP has Matern covariance with fixed smoothness parameter

Statistical approaches: Space-time

Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.94

Spatio-temporal change of support with application to American Community Survey multi-year period estimates

Jonathan R. Bradley*, Christopher K. Wikle and Scott H. Holan

Received 21 August 2015; Accepted 16 September 2015

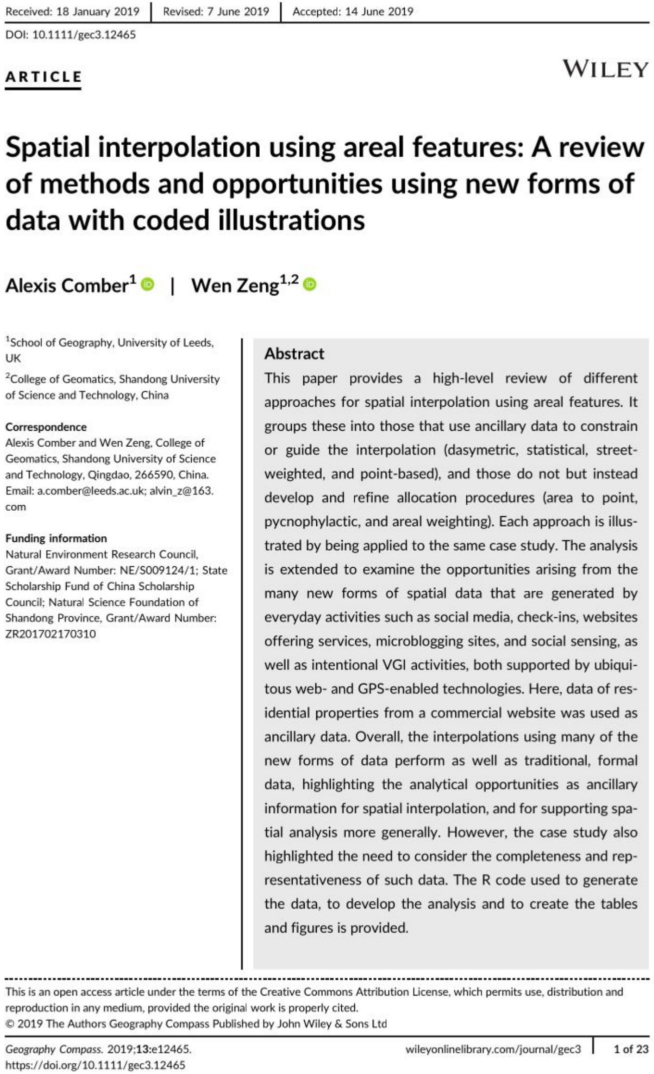
We present hierarchical Bayesian methodology to perform spatio-temporal change of support (COS) for survey data with Gaussian sampling errors. This methodology is motivated by the American Community Survey (ACS), which is an ongoing survey administered by the US Census Bureau that provides timely information on several key demographic variables. The ACS has published 1-year, 3-year, and 5-year period estimates, and margins of errors, for demographic and socio-economic variables recorded over predefined geographies. The spatio-temporal COS methodology considered here provides data users with a way to estimate ACS variables on customized geographies and time periods while accounting for sampling errors. Additionally, 3-year ACS period estimates are to be discontinued, and this methodology can provide predictions of ACS variables for 3-year periods given the available period estimates. The methodology is based on a spatio-temporal mixed-effects model with a low-dimensional spatio-temporal basis function representation, which provides multi-resolution estimates through basis function aggregation in space and time. This methodology includes a novel parameterization that uses a target dynamical process and recently proposed parsimonious Moran's I propagator structures. Our approach is demonstrated through two applications using public-use ACS estimates and is shown to produce good predictions on a hold-out set of 3-year period estimates.

Keywords: Bayesian; change-of-support; dynamical; hierarchical models; mixed-effects model; Moran's I; multi-year period estimate

- `stcos` CRAN package
- Gaussian errors
- non-stationary, non-separable, and asymmetric covariances

Other approaches

- Rich literature review, many w. use cases
- Mostly geography oriented
- Dasymetric mapping approaches



Received: 18 January 2019 | Revised: 7 June 2019 | Accepted: 14 June 2019
DOI: 10.1111/gec3.12465

WILEY

ARTICLE

Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations

Alexis Comber[1] | Wen Zeng[1,2]

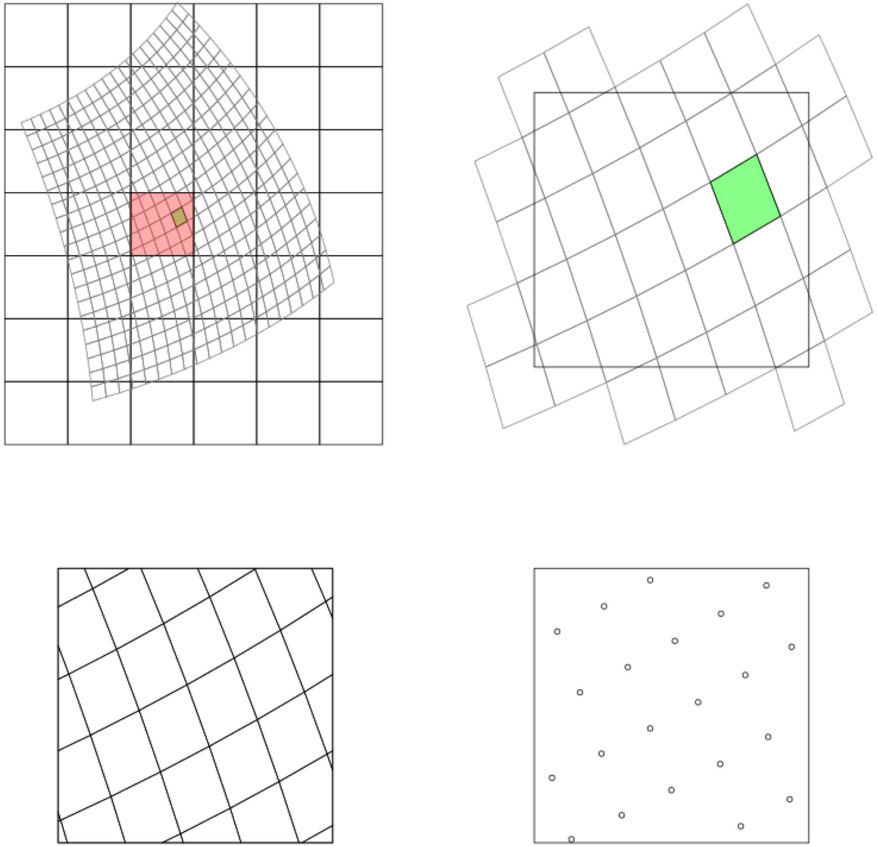
[1]School of Geography, University of Leeds, UK
[2]College of Geomatics, Shandong University of Science and Technology, China

Correspondence
Alexis Comber and Wen Zeng, College of Geomatics, Shandong University of Science and Technology, Qingdao, 266590, China.
Email: a.comber@leeds.ac.uk; alvin_z@163.com

Funding information
Natural Environment Research Council, Grant/Award Number: NE/S009124/1; State Scholarship Fund of China Scholarship Council; Natural Science Foundation of Shandong Province, Grant/Award Number: ZR201702170310

Abstract
This paper provides a high-level review of different approaches for spatial interpolation using areal features. It groups these into those that use ancillary data to constrain or guide the interpolation (dasymetric, statistical, street-weighted, and point-based), and those do not but instead develop and refine allocation procedures (area to point, pycnophylactic, and areal weighting). Each approach is illustrated by being applied to the same case study. The analysis is extended to examine the opportunities arising from the many new forms of spatial data that are generated by everyday activities such as social media, check-ins, websites offering services, microblogging sites, and social sensing, as well as intentional VGI activities, both supported by ubiquitous web- and GPS-enabled technologies. Here, data of residential properties from a commercial website was used as ancillary data. Overall, the interpolations using many of the new forms of data perform as well as traditional, formal data, highlighting the analytical opportunities as ancillary information for spatial interpolation, and for supporting spatial analysis more generally. However, the case study also highlighted the need to consider the completeness and representativeness of such data. The R code used to generate the data, to develop the analysis and to create the tables and figures is provided.

Raster-raster warping



Meng Lu, Marius Appel, Edzer Pebesma, 2018. Multidimensional Arrays for Analysing Geoscientific Data. ISPRS Int. J. Geo-Inf. 2018, 7(8)

Concluding remarks

- Most spatial data variables have block support, making down-scaling hard
- For up- and down-scaling, it matters whether variables are spatially extensive or intensive; this intel is typically not in our data sources
- Are covariates available at high (target) resolution? If yes:
  - Dasymetric mapping
  - DM + regression
  - Hierarchical modelling (INLA, TMB)
- For raster-raster:
  - use `gdalwarp`?
  - or exactextractr (`stars`/`terra`)?

gdalwarp:

-r <resampling_method>

Resampling method to use. Available methods are:

`near` : nearest neighbour resampling (default for non-COG drivers, fastest algorithm, worst interpolation quality).

`bilinear` : bilinear resampling.

`cubic` : cubic resampling (default for COG driver).

`cubicspline` : cubic spline resampling.

`lanczos` : Lanczos windowed sinc resampling.

`average` : average resampling, computes the weighted average of all non-NODATA contributing pixels.

`rms` root mean square / quadratic mean of all non-NODATA contributing pixels (GDAL >= 3.3)

`mode` : mode resampling, selects the value which appears most often of all the sampled points. In the case of ties, the first value identified as the mode will be selected.

`max` : maximum resampling, selects the maximum value from all non-NODATA contributing pixels.

`min` : minimum resampling, selects the minimum value from all non-NODATA contributing pixels.

`med` : median resampling, selects the median value of all non-NODATA contributing pixels.

`q1` : first quartile resampling, selects the first quartile value of all non-NODATA contributing pixels.

`q3` : third quartile resampling, selects the third quartile value of all non-NODATA contributing pixels.

`sum` : compute the weighted sum of all non-NODATA contributing pixels (since GDAL 3.1)