Introduction

Population data are often available at various levels of aggregation, such as census tracts or block groups in the United States or output areas in the United Kingdom. These units are often drawn for convenience and not because they represent some commonly accepted area in residents’ lived experiences. Despite this, researchers often use them as proxies for neighborhoods since they come with readily available demographic data. Researchers who do wish to move past these proxies and produce population estimates for other geographies typically resort to developing their own, often manual, implementations of the areal weighted interpolation workflow (Qiu, Zhang, & Zhou, 2012). The lack of reliable and easy to use software that implements this technique has therefore served as an additional barrier to using local neighborhood boundaries in research.

Areal weighted interpolation is the most accessible technique for estimating population values for overlapping but incongruent areas, such as imputing neighborhood populations from census tracts. Unlike interpolation techniques, such as inverse distance weighted interpolation, that are appropriate for point data, areal weighted interpolation is designed specifically for working with data already aggregated to some set of polygon spatial features (N. S.-N. Lam, 1983). Estimating population values from these data to an overlapping but incongruent set of polygon features is known as the polygon overlay problem (Goodchild, 1978), with the original data known as the “source” data and the overlapping set of features known as the “target” data (Markoff & Shapiro, 1973). The estimation process varies based on whether data are spatially extensive (i.e. count data) or spatially intensive (i.e. ratios, percentages, means, or medians; Goodchild & Lam, 1980).

In estimating values for the target features, areal weighted interpolation makes a significant assumption that individuals are evenly spread throughout the source features (Qiu et al., 2012). This assumption often breaks down in practice. For example, a census tract with a large park or a neighborhood that has a significant commercial area alongside residential housing would violate this assumption. While more complex methods do exist to compensate for violating this assumption, there are few tools to implement them, and the strategies have mainly remained the focus of academics instead of GIS practitioners (Langford, 2007; Qiu et al., 2012).

The areal R package

The areal package aims to reduce the barriers to the implementation of areal interpolation in spatial researchers’ work. By implementing the process in R, we aim to improve the reproducibility of the interpolation process, which is often done manually using point-and-click desktop GIS software. We also aim to provide additional functionality that is not currently available in existing approaches found in the sf package (Pebesma, 2018), for
example the `st_interpolate_aw()` function, while providing support for modern data
management (e.g. `tidyverse`) and spatial data (e.g. `sf`) frameworks.

Therefore, we provide functions that:

1. validate data suitability for areal interpolation,
2. allow for frictionless, iterative interpolation of both spatially extensive (e.g., count)
   and intensive (e.g., ratio) data, and
3. provide a manual workflow for implementing the interpolation process as well as a
   single omnibus function.

The validation process checks for five conditions that must be met prior to interpolation:

1. Are both objects `sf` objects?
2. Do both objects share the same coordinate system?
3. Is that coordinate system planar?
4. Do the given variables exist in the source data?
5. Will interpolation overwrite columns in the target data?

This validation process can be run independently by end users via the `ar_validate()`
function and is also run initially when `aw_interpolate()` is called. The validation process
is aimed at helping users new to both R and areal interpolation ensure they have arranged
their data correctly.

Unlike existing approaches, the `areal` package offers the ability to interpolate multiple
variables in a single function call and provides a wider set of options for making these
estimates. For spatially extensive interpolations, we offer a formula that matches the
existing functionality in `sf`, which is based on the total area of the original source feature
(specified with `weight = "total"`). Alternatively, we offer a second formula that is
more suitable for data where there should be complete overlap between the source and
target data, but there is not. Such incongruity could be due to data quality issues or
variation in how different sources represent particular geographies. This uses a sum of
the source feature areas remaining after the data are intersected as part of the spatial
weight calculation process (specified with `weight = "sum"`).

Our package also provides a manual approach for calculating estimates alongside the
primary function `aw_interpolate()`. This is important for users who need to unpack
the interpolation workflow and diagnose data issues that occur during interpolation as
well as programmers who want to use a portion of the workflow in their software. Finally,
`areal` provides both spatial (`sf` object) and a-spatial (`tibble` object) options for output.

A Short Example

Once data have been prepared for interpolation, meaning they share the same projected
coordinate system and there are no conflicts between variable names, interpolations can
be calculated with a single function `aw_interpolate()`. This functionality is illustrated
here using source and target data built into the `areal` package:
> `aw_interpolate(ar_stl_wards, tid = WARD, source = ar_stl_race,
+     sid = GEOID, weight = "sum", output = "tibble",
+     extensive = c("TOTAL_E", "WHITE_E", "BLACK_E"))`

# A tibble: 28 x 6

<table>
<thead>
<tr>
<th>OBJECTID</th>
<th>WARD</th>
<th>AREA</th>
<th>TOTAL_E</th>
<th>WHITE_E</th>
<th>BLACK_E</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;dbl&gt;</td>
<td>&lt;int&gt;</td>
<td>&lt;dbl&gt;</td>
<td>&lt;dbl&gt;</td>
<td>&lt;dbl&gt;</td>
<td>&lt;dbl&gt;</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>46138761</td>
<td>7992</td>
<td>153</td>
<td>7779</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>267817711</td>
<td>12145</td>
<td>1323</td>
<td>10639</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>66291644</td>
<td>7344</td>
<td>591</td>
<td>6635</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>53210707</td>
<td>8458</td>
<td>160</td>
<td>8203</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>60462396</td>
<td>8783</td>
<td>1526</td>
<td>7056</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>64337271</td>
<td>14050</td>
<td>5840</td>
<td>7439</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>101268146</td>
<td>15840</td>
<td>8220</td>
<td>6629</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>45966410</td>
<td>12188</td>
<td>7604</td>
<td>3796</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>73993891</td>
<td>14217</td>
<td>6838</td>
<td>6413</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>62915358</td>
<td>11239</td>
<td>8703</td>
<td>1667</td>
</tr>
</tbody>
</table>

# ... with 18 more rows

If a `sf` object is desired, changing the `output` argument in the example above to "sf" will provide not only the estimated values but the relevant geometric data as well.

Spatially intensive interpolations, which are used by analysts for data that are ratios or percentage values, can be calculated by replacing the `extensive` argument with `intensive`. Alternatively, both arguments and corresponding vectors of variable names can be provided to interpolate both spatially extensive and intensive data simultaneously.

The source data from the above data along with one of the variables, total population (`TOTAL_E`), are mapped in panel A below at the census tract level. The resulting data for the total population variable interpolated to the target political ward data have been mapped in panel B:

Additional examples and vignette illustrations of `areal`'s functionality are available on the package's website.
A Quick Comparison with sf

Since sf offers similar functionality, three comparisons are made and presented here. Code for these comparisons is available in an appendix within the areal GitHub repository. The first compares spatially extensive interpolations calculated in both areal and sf. The first ten features are shown here for comparison:

Comparison of sf and areal Output
Spatially Extensive Interpolation

<table>
<thead>
<tr>
<th>Ward</th>
<th>sf</th>
<th>areal, total weight</th>
<th>areal, sum weight</th>
<th>delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7990.572</td>
<td>7990.572</td>
<td>7991.565</td>
<td>-0.993</td>
</tr>
<tr>
<td>2</td>
<td>12041.971</td>
<td>12041.971</td>
<td>12145.021</td>
<td>-103.05</td>
</tr>
<tr>
<td>3</td>
<td>7333.783</td>
<td>7333.783</td>
<td>7344.287</td>
<td>-10.504</td>
</tr>
<tr>
<td>4</td>
<td>8457.676</td>
<td>8457.676</td>
<td>8457.672</td>
<td>0.004</td>
</tr>
<tr>
<td>5</td>
<td>8688.666</td>
<td>8688.666</td>
<td>8783.377</td>
<td>-94.711</td>
</tr>
<tr>
<td>6</td>
<td>14022.485</td>
<td>14022.485</td>
<td>14050.399</td>
<td>-27.915</td>
</tr>
<tr>
<td>7</td>
<td>15645.336</td>
<td>15645.336</td>
<td>15840.086</td>
<td>-194.75</td>
</tr>
<tr>
<td>8</td>
<td>12188.128</td>
<td>12188.128</td>
<td>12188.131</td>
<td>-0.002</td>
</tr>
<tr>
<td>9</td>
<td>14094.814</td>
<td>14094.814</td>
<td>14217.149</td>
<td>-122.335</td>
</tr>
<tr>
<td>10</td>
<td>11239.214</td>
<td>11239.214</td>
<td>11239.213</td>
<td>0</td>
</tr>
</tbody>
</table>

Both st_interpolate_aw() and aw_interpolate() with the weight = "total" produce identical results, which is expected.

The above comparison also shows the difference between using "total" and "sum" for the weight argument in aw_interpolate(). If we expect that our source and target data should overlap completely (but perhaps do not because of data quality issues), the "sum" approach will allocate all individuals into target features whereas "total" will not, instead allocating only a proportion of individuals relative to the overlap between the source and target features. In the example data provided in the areal package, this is the case - the Census tracts do not perfectly map onto the extent of the wards, and so the "total" approach is inappropriate.

The differences are highlighted in the delta column in the above figure. With these data, the "total" approach yields generally smaller results that the "sum" approach, with an average difference of -30.781. The largest difference between the "total" and "sum" approaches was -194.750 individuals. The added functionality within areal therefore provides a potentially more accurate set of estimations in similar cases where data should overlap, but do not.

The second comparison is with spatially intensive interpolations. For this approach, both st_interpolate_aw() and aw_interpolate() should yield identical results. As before, the first ten features are shown here for comparison:
As with the extensive interpolations where weight = "total", both `st_interpolate_aw()` and `aw_interpolate()` produce the expected identical results.

### Notes on Performance

Finally, a comparison of the speed at which both the sf and areal packages calculated these values is included in the appendix as well. For the extensive interpolations, `aw_interpolate()` performed 28 milliseconds slower on the sample data provided in the package, with a mean execution time of 279.429 milliseconds. Similarly, the intensive interpolations were 20 milliseconds slower when using areal, which had a mean execution time of 267.878 milliseconds. Both comparisons were made using the R package `microbenchmark` (Mersmann, 2018).

One particular performance concern to note are intersections (a step in the interpolation process) that return geometry collections. These will cause `st_interpolate_aw()` function from sf to error. The `aw_interpolate()` function will not error, but instead will correctly address these geometry collections before proceeding with the interpolation. However, the correction process can take significantly longer than it would a similarly sized data set that does not require this additional step.

The appendix also contains an example of interpolating population from Missouri’s 115 counties into its 4,506 Census block groups. When evaluated with `microbenchmark`, the average length of time to complete this process was 17.689 seconds.

### Availability

areal is open source software made available under the GNU GPL-3 license. It can be installed through CRAN (Prener & Revord, 2019) using: `install.packages("areal").` areal can also be installed from its GitHub repository using the remotes package: `remotes::install_github("slu-openGIS/areal").`
Acknowledgements

The authors wish to thank their colleagues, J.S. Onésimo Sandoval, Ph.D. and Taylor Harris Braswell, M.A., for reviewing an initial draft of this manuscript.

References


