Remaining gaps in open source software for Big Spatial Data

The volume and coverage of spatial data has increased dramatically in recent years, with Earth observation programmes producing dozens of GB of data on a daily basis. The term Big Spatial Data is now applied to data sets that impose real challenges to researchers and practitioners alike. The difficulties are partly related to a lack of tools supporting appropriate Coordinate Reference Systems (CRS). As rule, these data are provided in highly irregular geodesic grids, defined along equal intervals of latitude and longitude. Compounding the problem, users of such data end up taking geodesic coordinates in these grids as a Cartesian system, implicitly applying Marinus of Tyre's projection. A first approach towards the compactness of global geo-spatial data is to work in a Cartesian system produced by an equal-area projection. There are a good number to choose from, but those commonly supported by GIS software invariably relate to the sinusoidal or pseudo-cylindrical families, that impose important distortions of shape and distance. The land masses of Antarctica, Alaska, Canada, Greenland and Russia are particularly distorted with such projections. A more effective approach is to store and work with data in modern cartographic projections, in particular those defined with the Platonic and Archimedean solids. In spite of various attempts at open source software supporting these projections, in practice they remain today largely out of reach to GIS practitioners. This communication reviews persisting difficulties in working with worldwide big spatial data, current strategies to address such difficulties, the compromises they impose and the remaining gaps in open source software.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27215v1 | CC BY 4.0 Open Access | rec: 16 Sep 2018, publ: 16 Sep 2018

Remaining gaps in open source software for Big Spatial Data

Luís M. de Sousa[1]

[1]**ISRIC - World Soil Information,**
Droevendaalsesteeg 3, Building 101 6708 PB Wageningen, The Netherlands

Corresponding author:
Luís M. de Sousa[1]

Email address: luis.de.sousa@protonmail.ch

ABSTRACT

The volume and coverage of spatial data has increased dramatically in recent years, with Earth observation programmes producing dozens of GB of data on a daily basis. The term Big Spatial Data is now applied to data sets that impose real challenges to researchers and practitioners alike. The difficulties are partly related to a lack of tools supporting appropriate Coordinate Reference Systems (CRS). As rule, these data are provided in highly irregular geodesic grids, defined along equal intervals of latitude and longitude. Compounding the problem, users of such data end up taking geodesic coordinates in these grids as a Cartesian system, implicitly applying Marinus of Tyre's projection.

A first approach towards the compactness of global geo-spatial data is to work in a Cartesian system produced by an equal-area projection. There are a good number to choose from, but those commonly supported by GIS software invariably relate to the sinusoidal or pseudo-cylindrical families, that impose important distortions of shape and distance. The land masses of Antarctica, Alaska, Canada, Greenland and Russia are particularly distorted with such projections. A more effective approach is to store and work with data in modern cartographic projections, in particular those defined with the Platonic and Archimedean solids. In spite of various attempts at open source software supporting these projections, in practice they remain today largely out of reach to GIS practitioners. This communication reviews persisting difficulties in working with worldwide big spatial data, current strategies to address such difficulties, the compromises they impose and the remaining gaps in open source software.

1 INTRODUCTION

Raster datasets covering the entire globe are becoming ever more available, not only in the form of remote sensing derived products, but also as time-series of natural variables, such as those reporting to Climate, Geology or Sociology. Researchers in Geo-Informatics and Earth Sciences in general are thus increasingly able of working at the global scale. Remarkably, such datasets are in almost all cases provided in highly irregular global grids, defined along regular intervals longitude and latitude. Exemplary datasets include:

- Moderate Resolution Imaging Spectroradiometer (MODIS) - equal intervals of 0.05° latitude and longitude.
- Shuttle Radar Topography Mission (SRTM) - equal intervals of 1 arc second.
- Digital Terrain Elevation Data (DTED) - various levels with latitude intervals ranging from 1 to 30 arc seconds and longitude ranging from 1 to 180 arc seconds.
- IPCC climate scenarios - equal intervals of 2° latitude and longitude.

Researchers working at the global scale tend to use these data "as is", skipping any formal cartographic projection. Since any common GIS programme operates on the Cartesian plane, researchers end up tacitly working on the plane created by Marinus of Tyre's projection (in which the irregular global grid becomes a regular quadrangular grid). It is deeply ironic that Earth Sciences continue relying on a mathematical formulation that is almost 2 000 years old.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27215v1 | CC BY 4.0 Open Access | rec: 16 Sep 2018, publ: 16 Sep 2018

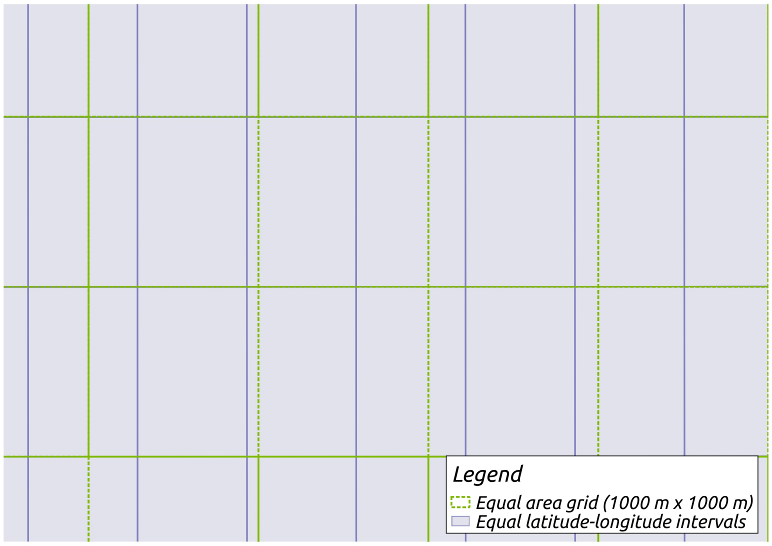

Figure 1. A regular grid defined on an equal-area projection compared with a grid defined on equal intervals of latitude and longitude, at a latitude of 50°.

While the area of the Earth's surface is in the order of 510 Mm^2, the total area of Cartesian space created by Marinus of Tyre's projection is over 800 Mm^2, a difference of 60%. This also means that a global dataset sampling the Earth at regular angular intervals, contains 60% more samples than one defined to favour regularity of sampling areas (Figure 1 exemplifies this difference). This is not only a problem for storage space, but much more so to Geo-computation with big data, demanding more memory and computing cycles. When employing modern techniques such as Neuronal Networks or Machine Learning, researchers can easily be facing computation constrains with Marinus of Tyre's projection that in an equal area projection would be less restrictive or altogether non existent.

A number of reasons collude to deter researchers from working with alternative cartographic projections:

- Other datasets are also provided in similar constant latitude-longitude global grids.
- Re-projecting original data with an alternative cartographic projection can be computationally expensive.
- Re-projection may lead to data loss.
- It is not easy to identify the most appropriate Cartographic projection.

The last item is itself rooted in another issue: support for modern equal-area cartographic projections remains scant in free and open source software for Geo-Informatics (FOSS4G). Only those equal-area projections yielding higher shape distortions are readily usable in stock open-source GIS programmes, which naturally plays against their adoption.

This article starts with a brief review of the issues with popular equal-area projections (Section 2) it then reviews current support to modern projections of this class in FOSS4G (Section 3). Section 4 concludes by identifying the development avenues in FOSS4G software enabling work with appropriate equal-area projections.

2 POPULAR EQUAL-AREA PROJECTIONS

The gains in storage space and the consequent reduction in computation demands well justify working with global rasters on equal-area projections. The options are many for the purpose, however, only a few of these projections are actually operational in an open source software stack. Figure 2 portraits schematically the dependencies of various off-the-shelf geo-spatial open source programmes. Essentially, any cartographic projection must be supported by both Proj (PROJ contributors, 2018) and GDAL (GDAL/OGR

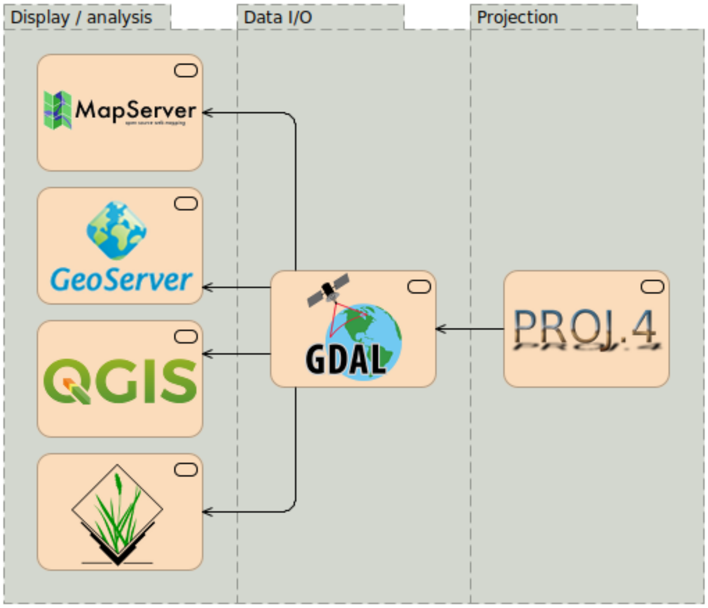

Figure 2. Dependencies of staple FOSS4G programmes.

contributors, 2018) to be fully usable in an open source environment. Unfortunately, the slim number of supported equal-area projections are mostly part of the sinusoidal or pseudo-cylindrical families, inducing deep shape distortions. In large measure this is due to GDAL, that requires the inverse of any projection that it supports.

Figure 3 provides a practical example with the portrayal of New Zealand in a projection centred on 0° E, 0° N. The islands composing this country turn out deeply warped, particularly soo with the classical Sinusoidal projection. This is significant, given how popular this last projection is (Seong et al., 2002), used even by institutions like NASA (possibly because it preserves distances along parallels and the central meridian). Such distortions are particularly critical for datasets that while produced on a global scale, are relevant to local analysis process, e.g. climate, environment.

A detailed analysis of these distortions can be obtained using Tissot's Indicatrix (Goldberg and Gott III, 2007). However, Figure 3 is enough to show the distortions imposed by these popular projections. And it is not only a visual accuracy problem, shapes warped this much also imply higher information loss during re-projection to alternative CRSs, e.g. in a local analysis exercises. The need for different equal-projections is therefore well patent, so that space and computation economy does not come at the expense of cartographic quality.

A further alternative is to use multi-projection systems, usually one per continent, such as the Equi7 Bauer-Marschallinger et al. (2014), but these create problems of their own. There is an overhead in managing different projections simultaneously and overlaps between the various Cartesian spaces in the system can create problems of their own. They also make the publication of data to third parties cumbersome, since standards like WMS or WCS do not consider multi-projection systems.

3 MODERN EQUAL-AREA PROJECTIONS

Throughout the past century several mathematicians and cartographers produced novel equal-area projections that considerably ameliorate shape distortion. However, for one reason or other, none of them is fully supported in a FOSS4G stack. This section explores a few of them.

3.1 Hammer's

This projection was developed in the last decade of the XIX century, with the goal of ameilorating the distortions produced by Mollweide's projection at high longitudes (Snyder, 1997). Hammer drew inspiration from Aitoff's projection, using an elliptical counter-domain where all parallels are curved (apart from the one passing in the central point, usually the Equator).

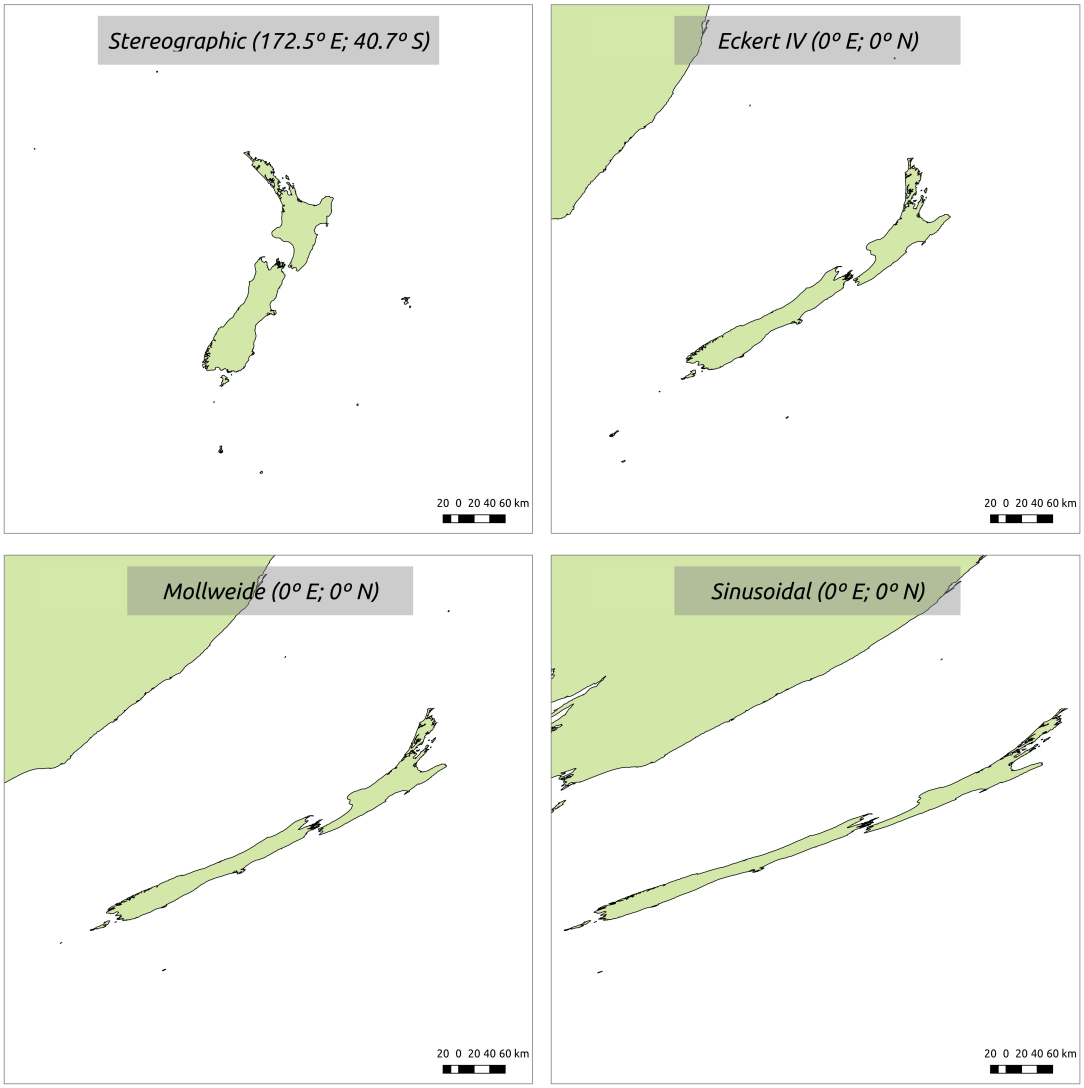

Figure 3. The distortions applied on New Zealand by three popular equal-area map projections when applied on the point with coordinates (0° E, 0° N).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27215v1 | CC BY 4.0 Open Access | rec: 16 Sep 2018, publ: 16 Sep 2018

In spite of its improvements over Mollweide's, Hammer's projection never found the popularity of the former. Perhaps for that reason, Hammer's projection is scantly supported by FOSS4G, only by Proj, in its direct form.

Listing 1. Hammer's projection with Proj and GDAL

```
$ cs2cs +init=epsg:4326 +to +proj=hammer +lat_0=0 +lon_0=0 +datum=WGS84
+units=m +no_defs <<EOF
> 15 -15
> EOF
1625591.45 -1668538.36 0.00

$ gdaltransform -s_srs "+init=epsg:4326" -t_srs "+proj=hammer␣+lat_0=0
+lon_0=0␣+datum=WGS84␣+units=m␣+no_defs" <<EOF
> 15 -15
> EOF
ERROR 1: Translating source or target SRS failed:
+proj=hammer +lat_0=0 +lon_0=0 +datum=WGS84 +units=m +no_defs
```

3.2 Goode's Homolosine

John P. Goode (1925) developed his Homolosine projection in the 1920s, while attempting to interrupt Mollewiede's projection. The end result was a major improvement over the then state-of-the-art of equal-area projections, with serious shape distortions only present at latitudes over 60°. This projection would gain popularity in the following decades, often used to convey socio-economic information; it is easy to find it in didactic and technical publications of the second half of the XX century. However, with the advent of web mapping, this projection has almost disappeared from general interest publications.

Interestingly, Goode's Homolosine is in fact reasonably supported by FOSS4G. Both Proj and GDAL fully support it, while analysis programmes like QGIS or GRASS are able to intake Homolosine rasters. Unfortunately, the analysis programmes are not able to correctly use and portray vector data encoded in this projection, greatly limiting analysis and cartography. Figure 4 presents raster and vector maps as portrayed by QGis in this projection; at least the vector rendering library as it a loss with Goode's interruptions. Even so, Goode's Homolosine is the closest it gets to a proper equal-area projection suitable for modern day big geo-spatial data processing.

3.3 Bogg's Eumorphic

Just a few years after Goode, Samuel W. Bogg developed a projection that produces a similarly shaped world map (Snyder, 1997). Bogg used an average of the Sinusoidal and Molleweide's projections to obtain easting coordinates and a pseudo-cylindrical to obtain the northing. A map produced with Bogg's Eumorphic projection can be easily mistaken by Goode's Homolosine. Perhaps for coming later, the Eumorphic never became as popular as the Homolosine.

Bogg's Eumorphic projection is supported by Proj, but not by GDAL, meaning it is in practice largely unusable with a FOSS4G stack. Rending it operational would require at least the implementation of its inverse in Proj, so that GDAL can accept it. Beyond that, the same issues with vector data in the Homolosine are to be expected with the Eumorphic.

Listing 2. Bogg's projection with Proj and GDAL

```
$ cs2cs +init=epsg:4326 +to +proj=boggs +lat_0=0 +lon_0=0 +datum=WGS84
+units=m +no_defs <<EOF
> 15 -15
> EOF
1541082.76      -1755739.47 0.00

$ gdaltransform -s_srs "+init=epsg:4326" -t_srs "+proj=boggs␣+lat_0=0
+lon_0=0␣+datum=WGS84␣+units=m␣+no_defs" <<EOF
> 15 -15
> EOF
ERROR 1: Translating source or target SRS failed:
+proj=boggs +lat_0=0 +lon_0=0 +datum=WGS84 +units=m +no_defs
```

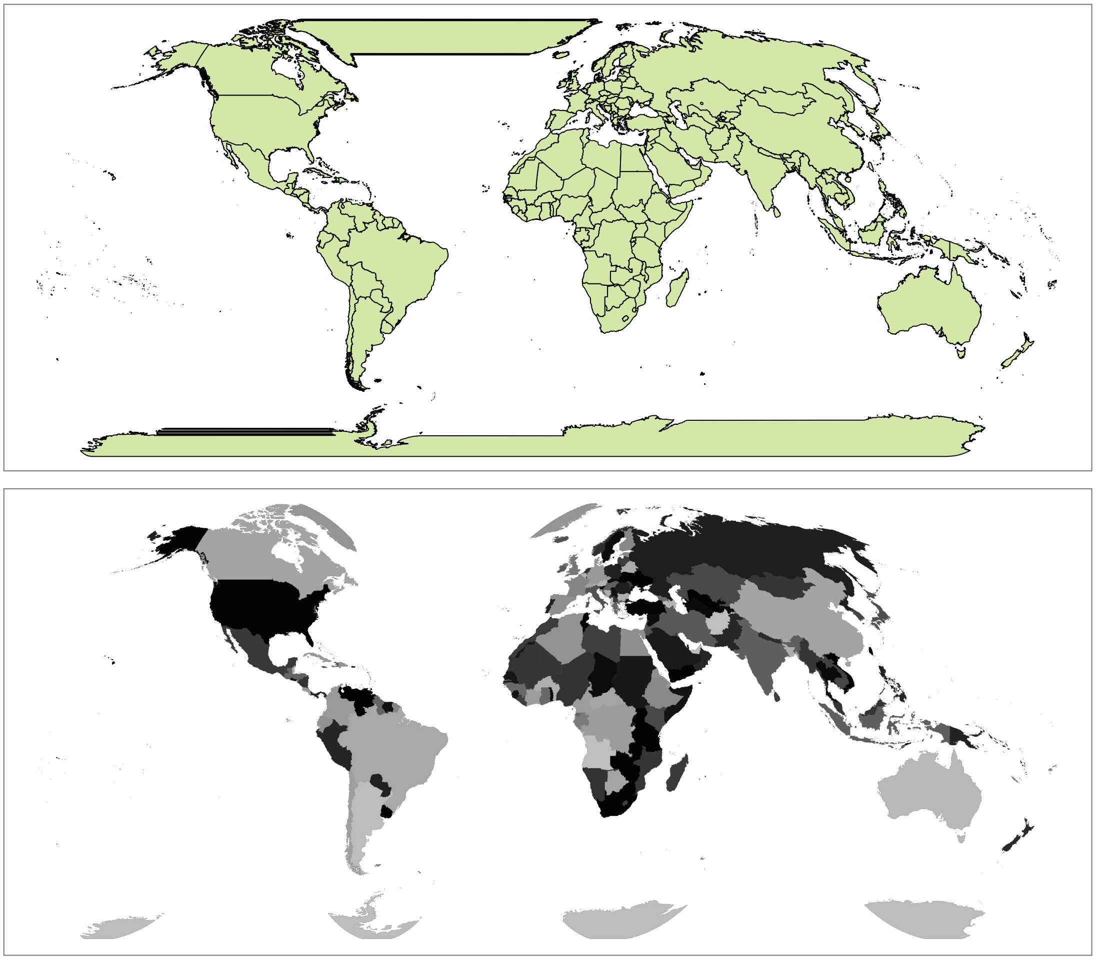

Figure 4. World maps portrayed by QGis with Goode's projection, vector top, raster bottom.

3.4 Snyder's Icosahedral

John P. Snyder (1992) produced a number of reference scholarly works on Cartography in the last decades of the XX century and developed his own projections. Most notable is the equal-area projection Snyder developed for Archimedian and Platonic solids. Fuller (1943) had earlier used the icosahedron in his somewhat famous Dymaxion projection, but this last author employed rather a conformal projection. Snyder went further, with an equal-area formulation, that is extendable to the dodecahedron and truncated icosahedron.

Snyder's Icosahedral projection was immediately picked up by researchers working with the Environmental Protection Agency of US, whom at the time were developing global sampling grids on the geodetical domain. Kevin Sahr would produce an open source programme that creates global geodetical grids, in the process implementing Snyder's Icosahedral projection (Sahr et al., 2003). This programme was left dusting for decades until recently, when Barnes et al. (2017) created a binding software package that allows the usage of Sahr's software with the R programming language.

The Icosahedral equal-area projection is partially supported by Proj, limited to two pre-defined orientations of the icosahedron relative to the globe. No other FOSS4G supports this projection, making it impractical for off-the-shelf spatial analysis.

Listing 3. Snyder's Icosahedral projection with Proj and GDAL

```
$ cs2cs +init=epsg:4326 +to +proj=isea +lat_0=0 +lon_0=0 +datum=WGS84
+units=m +no_defs <<EOF
> 15 -15
> EOF
6659048.10      7609090.35 0.00

$ gdaltransform -s_srs "+init=epsg:4326" -t_srs "+proj=isea␣+lat_0=0
+lon_0=0␣+datum=WGS84␣+units=m␣+no_defs" <<EOF
> 15 -15
> EOF
ERROR 1: Translating source or target SRS failed:
+proj=isea +lat_0=0 +lon_0=0 +datum=WGS84 +units=m +no_defs
```

3.5 Snyder's Dodecahedral and Truncated-Icosahedral

Snyder's projection for the icosahedron is directly applicable to the dodecahedron and the truncated-icosahedron. This last solid is built from the icosahedron, adding a pentagonal face in place of each of its twelve vertices; the result is a solid with 32 faces, 20 hexagonal and 12 pentagonal. The Truncated-Icosahedral is actually a composite, with two different projections, at two different scales, one for the hexagonal faces and another for the pentagonal.

According to Snyder's calculations, both Dodecahedral and the Truncated-Icosahedral projections yield lower shape distortions rates than the Icosahedral and the author himself appeared to favour these. Moreover, the representation of continents in the Dodecahedral projection is visually more palatable. However, no practical open source implementations of these projections could be identified.

4 SUMMARY AND CONCLUSIONS

This article reviewed the inadequacy of popular cartographic projections for the analysis of big spatial data, particularly in raster form. The most popular of all, Marinus of Tyre's equirectangular projection, induces irregular grids that needlessly expand the number of cells in rasters, with penalties in storage space and computation load. Equal-area projections address this, however, those supported by FOSS4G are scant and invariably impose deep shape distortions that lead to problems of their own.

Support for modern equal-area projections is thus an obvious necessity for the FOSS4G community. Only Goode's projection is presently an option, and strictly regarding raster data. Considering the tests reported above, four different development pathways can be devised in this field:

a) reconsider the requirement for inverse projections, wherever practical;

b) develop missing inverse projections in the Proj package;

c) implement further polyhedral projections in Proj;

d) expand support to equal-area projections in vector portrayal and processing libraries.

One wonders what Marinus of Tyre would think, were he to know how popular his projection remains in the computer age. The father of mathematical cartography might ask himself why map projections in general are still so much in use, when parchments are no longer necessary to store and present geo-spatial data. In effect, modern equal-area projections play a capital role in upgrading geo-computation to the Geodetical domain (Sahr et al., 2003), justifying in a further way the investment from the community.

REFERENCES

Barnes, R., Sahr, K., Evenden, G., Johnson, A., and Warmerdam, F. (2017). dggridr: discrete global grids for r.

Bauer-Marschallinger, B., Sabel, D., and Wagner, W. (2014). Optimisation of global grids for high-resolution remote sensing data. *Computers & Geosciences*, 72:84–93.

Fuller, R. B. (1943). *Life Presents R. Buckminster Fuller's Dymaxion World*. Life.

GDAL/OGR contributors (2018). *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation.

Goldberg, D. M. and Gott III, J. R. (2007). Flexion and skewness in map projections of the earth. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 42(4):297–318.

Goode, J. P. (1925). The homolosine projection: a new device for portraying the earth's surface entire. *Annals of the Association of American Geographers*, 15(3):119–125.

PROJ contributors (2018). *PROJ coordinate transformation software library*. Open Source Geospatial Foundation.

Sahr, K., White, D., and Kimerling, A. J. (2003). Geodesic Discrete Global Grid Systems. *Cartography and Geographic Information Science*, 30(2):121–134.

Seong, J. C., Mulcahy, K. A., and Usery, E. L. (2002). The sinusoidal projection: A new importance in relation to global image data. *The Professional Geographer*, 54(2):218–225.

Snyder, J. P. (1992). An equal-area map projection for polyhedral globes. *Cartographica*, 29(1):10–21.

Snyder, J. P. (1997). *Flattening the earth: two thousand years of map projections*. University of Chicago Press.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27215v1 | CC BY 4.0 Open Access | rec: 16 Sep 2018, publ: 16 Sep 2018