

# SCIENTIFIC DATA

## OPEN Data Descriptor: Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010

Received: 23 February 2018

Accepted: 16 August 2018

Published: 30 October 2018

Marius Gilbert<sup>1,2</sup>, Gaëlle Nicolas<sup>1</sup>, Giusepina Cinardi<sup>3</sup>, Thomas P. Van Boeckel<sup>4,5</sup>, Sophie O. Vanwambeke<sup>6</sup>, G. R. William Wint<sup>7</sup> & Timothy P. Robinson<sup>3</sup>

Global data sets on the geographic distribution of livestock are essential for diverse applications in agricultural socio-economics, food security, environmental impact assessment and epidemiology. We present a new version of the Gridded Livestock of the World (GLW 3) database, reflecting the most recently compiled and harmonized subnational livestock distribution data for 2010. GLW 3 provides global population densities of cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in each land pixel at a spatial resolution of 0.083333 decimal degrees (approximately 10 km at the equator). They are accompanied by detailed metadata on the year, spatial resolution and source of the input census data. Two versions of each species distribution are produced. In the first version, livestock numbers are disaggregated within census polygons according to weights established by statistical models using high resolution spatial covariates (dasymetric weighting). In the second version, animal numbers are distributed homogeneously with equal densities within their census polygons (areal weighting) to provide spatial data layers free of any assumptions linking them to other spatial variables.

Design Type(s)	data integration objective • parallel group design • process-based data analysis objective
Measurement Type(s)	livestock abundance
Technology Type(s)	digital curation
Factor Type(s)	animal • geographic location
Sample Characteristic(s)	chicken • Earth (Planet) • anthropogenic habitat • cattle • Bovinae • Ovis aries • Capra aegagrus • Equus • Sus • Anatidae

<sup>1</sup>Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Brussels, Belgium. <sup>2</sup>Fonds National de la Recherche Scientifique (FNRS), Brussels, Belgium. <sup>3</sup>Animal Production and Health Division (AGA), Food and Agriculture Organization of the United Nations, Rome, Italy. <sup>4</sup>Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland. <sup>5</sup>Center for Diseases Dynamics Economics and Policy, Washington DC, USA. <sup>6</sup>Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium. <sup>7</sup>Environment Research Group Oxford (ERGO), Department of Zoology, Oxford, United Kingdom. Correspondence and requests for materials should be addressed to M.G. (email: mgilbert@ulb.ac.be) or T.P.R. (email: timothy.robinson@fao.org)

## Background & Summary

Livestock play a key role in global food systems as the main source of animal protein (milk, meat and eggs), contribute to crop productivity through the provision of draught power and manure, and to the livelihoods and nutrition of poor households in low- and middle-income countries<sup>1</sup> (LMICs). Livestock farming has a major impact on the environment, through greenhouse gas (GHG) emissions from enteric fermentation and manure, disruption of nitrogen and phosphorous cycles and indirect impacts on biodiversity and other ecosystem services through overgrazing and land-use change<sup>2</sup>. Livestock farming also bears public health implications through its role in food-borne disease transmission, the emergence and spread of infectious zoonotic diseases<sup>3</sup> such as avian influenza<sup>4</sup>, Q-fever and MERS and its contribution to the global burden of antimicrobial resistance, linked to the routine abuse of those drugs in livestock production.<sup>5,6</sup> Detailed, contemporary data sets on the global distribution of the most important species of farmed animals have a wide range of applications in understanding the social, economic, environmental, epidemiological and public health impacts of the livestock sector.

The gridded livestock of the world database (GLW 1) produced in 2007 had three objectives<sup>7</sup>: i) to collect, harmonize and disseminate subnational global livestock data, ii) to predict livestock numbers in areas with missing census counts (gap-filling), and iii) to provide a statistically-informed estimate of how livestock may be distributed within census units (downscaling). GLW 1 was produced at a spatial resolution of 0.0416666 decimal degrees (approximately 5 km at the equator). In 2014, an updated GLW 2 was published, benefiting from the availability of finer-scale and more contemporary input census data, from the improvement of the processing and from higher spatial resolution predictor variables that were used for downscaling<sup>8</sup>.

In this paper, we describe a new global, subnational livestock dataset (GLW 3) generated using Random Forests (RF), a machine-learning technique recently shown to provide more accurate gap-filling and disaggregation of livestock data than did the previously-used multivariate regression methods<sup>9</sup>. In addition to that important change in methodology, GLW 3 differs from the previous ones in three ways.

For each species, we now provide a detailed report that includes comprehensive metadata on the input census data for each country (e.g. year, resolution and source) and goodness-of-fit metrics of the models by continent and by the size of the administrative unit from which the census data came. This enables users to assess the quality of the estimates for each combination of species, country and size of census unit.

All species distributions are now available in two representations, termed dasymetric (DA) and areal-weighted (AW). The DA models correspond to previous GLW versions, whereby different animal densities are assigned to different pixels within a given census polygon according to the RF models. In contrast, the AW models simply spread individuals of a census polygon evenly, and the density of animals in each pixel corresponds to the average number of animals per km<sup>2</sup> of suitable land in the census unit. The AW models were introduced because the spatial predictor variables used in the downscaling algorithms (e.g. human population density, vegetation indices and topography) may introduce uncontrolled confounding effects or circularity for users wishing to study livestock distribution numbers independently of any other spatial variables. The AW models are free of the influence of other spatial predictor variables, at the cost of displaying cruder distribution patterns, especially in large census areas containing a wide range of different environmental, land-use and farming conditions. In polygons where input census data were missing, the AW model simply includes the aggregated predictions of the DA models, and a separate layer is provided for the user that distinguishes between predictions and census observations.

GLW 3 provides global data (DA, AW and prediction status) at a spatial resolution of 0.083333 decimal degrees (approximately 10 km at the equator), as the higher spatial resolution of previous GLW versions could be misleading in areas where the census data were of poor quality.

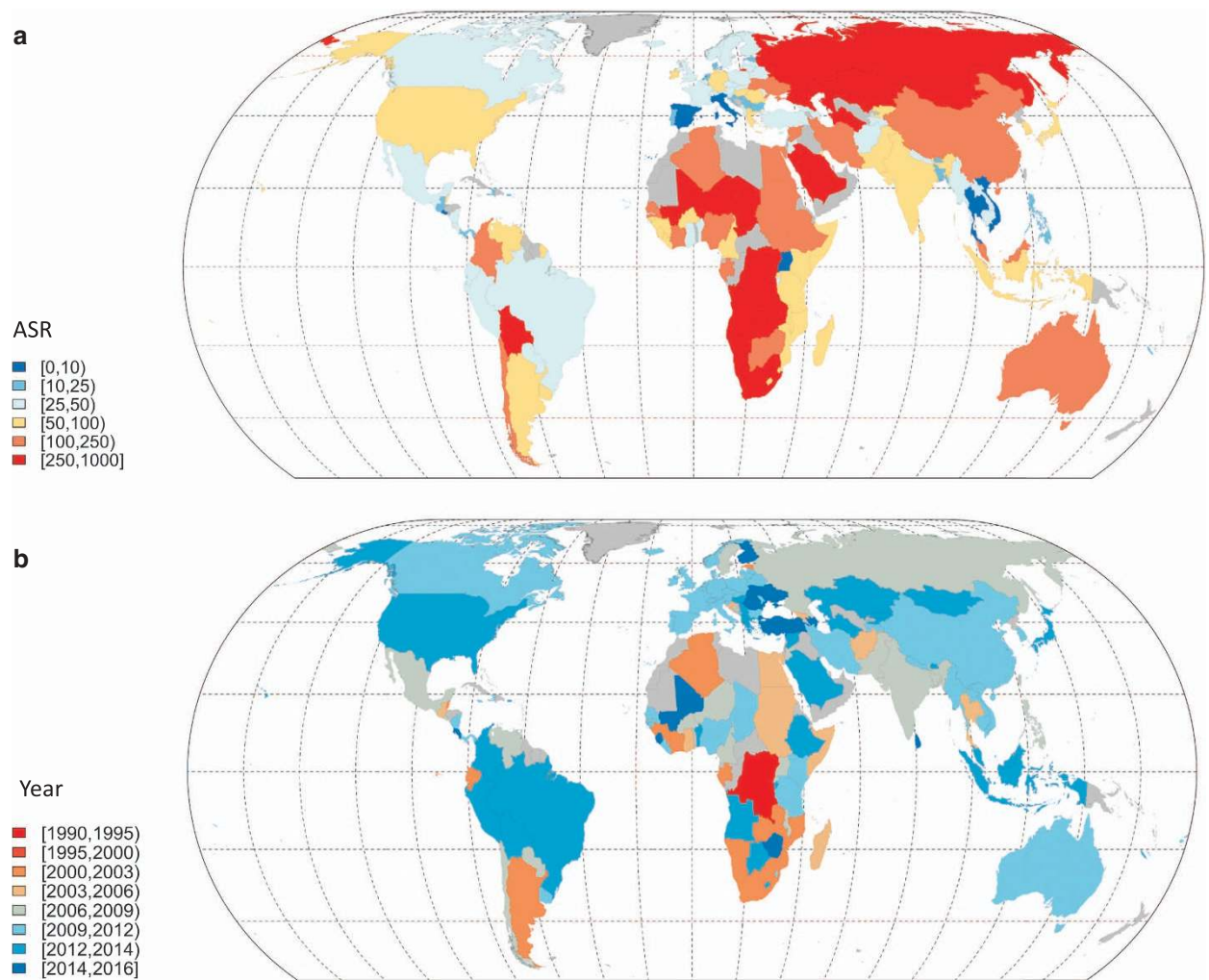
Future versions of GLW will differentiate stocks according to production systems for ruminant (meat vs. dairy) and monogastric species (intensive vs. extensive, meat vs. egg production). Higher resolution models for individual countries where the census data can support such predictions will also be produced.

## Methods

The only change to the overall workflow, which was fully detailed for GLW 2 in Robinson *et al.*<sup>8</sup>, is that in GLW 3 RF models have replaced stratified linear multiple regressions for predictions<sup>9</sup>.

## Data mining

Detailed livestock census statistics are mined from agricultural yearbooks or through direct contacts with ministries or statistical bureaux. The census statistics are usually found in the form of numbers per administrative unit, in which case they need to be linked to corresponding geographic information system (GIS) boundaries. Data are increasingly found though as pre-prepared GIS files that are then integrated in a centralized database. These individual country data are combined into a global database, which often implies resolving typology issues; miss-matched, split or merged polygons, for example. In compiling GIS data from subnational census counts priority is given to censuses that most closely match the reference year (2010 for GLW 3) and those with the highest level of spatial detail. This results in a global mosaic of data from different spatial resolutions and different years. For example, Fig. 1a illustrates the

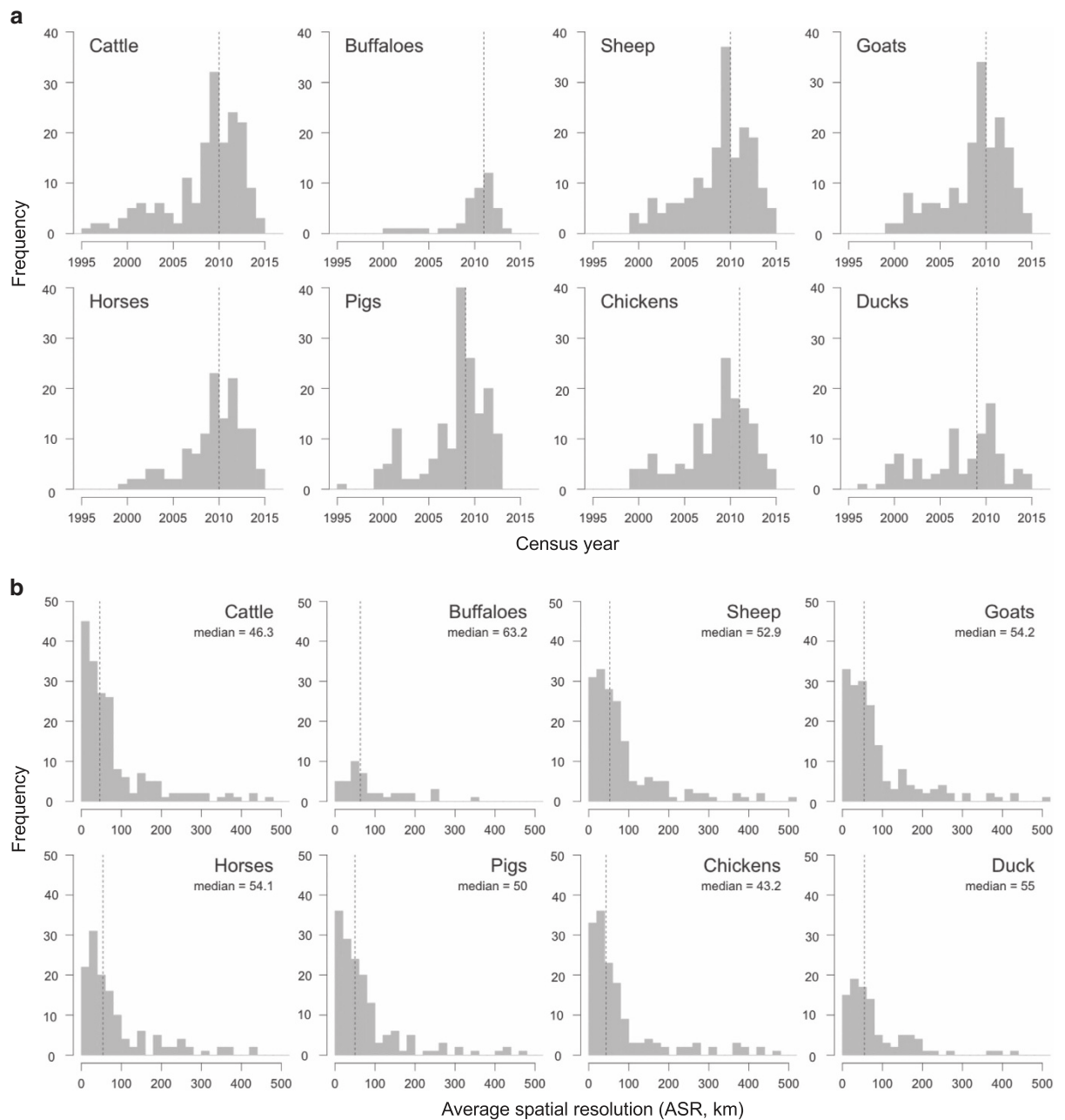


**Figure 1.** Maps of the GLW 3 average spatial resolution (ASR) and year of the chicken census data. Countries where no subnational census data could be obtained are indicated in grey.

heterogeneity in the Average Spatial Resolution (ASR) - the square root of the mean area of the census units - of the input data in each country for chickens. Figure 1b shows the year of each census. These figures highlight the large variability in ASR, with countries such as Italy and Thailand having very detailed subnational data (ASR < 10 km and a mean area < 100 km<sup>2</sup>), and countries such as Russia or South Africa with very coarse subnational census data (ASR > 250 km and mean area > 62 500 km<sup>2</sup>). There are also important differences in the census years. The oldest subnational census data are for the Democratic Republic of the Congo (1994), whilst some countries, for example Turkey, have very recent data (2014). Both the ASR and year of the census data depend on the species in question, as sometimes data can be available for one species and not for others. The two types of information are essential indicators of data quality and are therefore provided in the metadata. Summarized distributions of ASR and census years of GLW 3 are shown for all species in Fig. 2a and b, and are detailed in the metadata for each species.

#### Estimating densities corrected for unsuitable areas

Densities are estimated in each of the census polygon by dividing the number of animals from the census by the surface area of the administrative unit polygon (estimated in an Albert equal area projection), corrected by a mask excluding unsuitable areas. The suitability mask is very conservative and only excludes permanent water (pixels covered by >50 percent of water, see data source in Table 1), and areas where human population densities exceed 5,000 (North America, Europe and Oceania), 7,500 (South America) or 10 000 (Asia and Africa) people km<sup>-2</sup> as defined by the human population data layer (see data source in Table 1). Those different thresholds are used to account for the fact that urban population density is often higher in LMICs, where small-scale livestock farming may continue deeper into peri-urban and urban areas. The thresholds were conservatively defined to exclude only the core urban centres following an exploratory data analysis of human population density in urban pixels defined by the



**Figure 2.** Distribution of input census data year and average spatial resolution per country for the different livestock species. The median year of the subnational census (a) is indicated with the vertical dotted line. The median spatial resolution (b) ranges between 43.2 and 55 km, and shows considerable variability, with some countries having an ASR well above 100 km (corresponding to 10,000 km<sup>2</sup>).

MODIS global land cover 2010<sup>10</sup> or with >50% built up areas in the Global Human Settlement Layer of 2014<sup>11</sup>. In addition, a global mask of protected areas is derived from the 2010 version of the World Database on Protected Areas (Table 1). The International Union for the Conservation of Nature (IUCN) categories Ia and Ib, II, and III were masked as unsuitable as these are characterised by stringent conservation measures and tight regulation of human activity – the encroachment of roaming cattle and other grazing activities is therefore less likely in these than in other areas<sup>12</sup>.

#### Sampling and extraction of predictor covariates

Sampling points are distributed across the geographical space and the values for the suitability-corrected livestock densities are extracted from the subnational census data, constituting the dependent variable. The values of the predictor variables, listed in Table 1, are extracted for each of the sample pixels. All GIS raster layers of inputs (e.g. masks and predictor variables) and outputs (DA and AW predicted densities)

Type	Variables	Use	Source
Land	Land and water area	Spatial domain	CIESIN <sup>21</sup>
Land use	IUCN world database of protected area	Mask	UNEP-WCMC <sup>22</sup>
Anthropogenic	Human population density (consensus model between Worldpop, Landsan and GPW4)	Spatial predictor and suitability mask	Tatem <i>et al.</i> <sup>23</sup> , Dobson <i>et al.</i> <sup>24</sup> , CIESIN <sup>25</sup>
	Travel time to cities of 50,000 people	Spatial predictor	Nelson <i>et al.</i> <sup>26</sup>
Topography	Elevation (GTOPO30)	Spatial predictor	LDAAC <sup>27</sup>
	Slope (GTOPO30)	Spatial predictor	LDAAC <sup>27</sup>
Vegetation	10 Fourier-derived variables from Normalized Difference Vegetation Index from MODIS (MODIS)*	Spatial predictor	Scharlemann <i>et al.</i> <sup>28</sup>
	Length of growing period	Spatial predictor	Jones & Thornton <sup>29</sup>
	Green-up and senescence (annual cycle 1 and 2)	Spatial predictor	Zhang <i>et al.</i> <sup>30</sup>
	Cropping intensity	Spatial predictor	Fritz <i>et al.</i> <sup>31</sup>
	Forest cover	Spatial predictor	Hansen <i>et al.</i> <sup>32</sup>
Climatic	14 Fourier-derived variables from Day Land Surface Temperature (MODIS)	Spatial predictor	Scharlemann <i>et al.</i> <sup>28</sup>
	Precipitations	Spatial predictor	Hijmans <i>et al.</i> <sup>33</sup>

**Table 1. List of input spatial dataset used in the production of the GLW datasets.** \* Annual mean, annual minimum, annual maximum, amplitude and phase of annual cycle, amplitude and phase of bi-annual cycle, amplitude and phase of tri-annual cycle, variance in annual, bi-annual, and tri-annual cycles.

are processed with a global extent and a spatial resolution of 5 minutes of arc, i.e. 0.083333 decimal degrees, which corresponds approximately to 10 km at the equator. The sampling strategy is identical to that described in Robinson *et al.*<sup>8</sup>, i.e. balancing the sampling between the most detailed census data while ensuring a sufficient geographical coverage in areas with less detailed input data. As a result, a minimum of one sampling point is taken from each census polygon, and additional points are added proportionally to the polygon surface area with a sampling density of one point for every 10 000 km<sup>2</sup>.

### Random Forest models and cross-validation

The sample points are divided between training and validation sets according to the subnational census polygons, using sample points from 70 percent of the polygons for training the models and from 30 percent of the polygons for assessing the model accuracy. This operation is repeated 5 times, each time selecting a different set of polygons to train the RF models and goodness-of-fit (GOF) measurements. The parameters of the RF models were investigated in Nicolas *et al.*<sup>9</sup> and are set as follows: i) a third of the variables are used to build each tree with a minimum of 5 variables; ii) the number of trees is set to 1/20<sup>th</sup> of the number of sampling points, with a minimum of 100 trees; and iii) the node size is set to 1/1,000<sup>th</sup> of the number of sampling points, with a minimum node size of 5. The 5 bootstrapped RF models are then applied to the raster predictor variables to estimate a density value in each pixel, and the 5 predicted values are used to estimate the prediction mean and standard deviation in each pixel.

The 30 percent of polygon data that were held back for each bootstrap are used to estimate GOF metrics. For each of these polygons and each bootstrap, the predicted values of the pixels falling into the polygon are summed to calculate a predicted total per polygon. The predicted and observed animal numbers of these 30% validation polygons are then used to estimate the Root Mean Squared Error (RMSE) and correlation coefficient between the observed and predicted totals, as measures, respectively, of accuracy and precision (see technical validation section).

The entire analysis is stratified by developing 5 bootstrapped models for each continent: North America, South America, Europe, Africa, Asia and Oceania. In total 30 RF models are produced and applied to six continents. The GOF metrics are produced separately for each continent.

### Post-processing

For the dasymetric product, the average values predicted by the RF models are used as weights to distribute the animals within each subnational census unit at the pixel level. For each polygon, the pixel weights are multiplied by the ratio of the total number of animals per census unit to the sum of pixel weights in the polygon (i.e. total of the RF model mean prediction). In polygons with missing census values the RF mean predicted density is used (i.e. the factor applied to the weights = 1). In some countries, the spatial census units have smaller areas than the area of a single pixel (approx. 100 km<sup>2</sup> at the equator). In these cases, the sum of the animals from those small census polygons falling within a pixel is estimated and assigned to that pixel; replacing the RF model prediction. This can result in two situations. When pixels are smaller than their corresponding polygon, the sum of the pixel values matches the observed animal total of the polygon. Conversely, when pixels are larger than their corresponding polygons, the pixel value matches the sum of the intersecting smaller polygons. In both situations, issues of polygon boundaries going across pixel boundaries are resolved according to the proportion of the intersecting surface area.

Species	ASR median	Year min./median/max.	Spatial units	Density Median/99% percentile	Data citation
Cattle	46.3 km	1984/2010/2014	74,035	12.98/277.42	1
Buffaloes	63.2 km	2002/2011/2013	31,900	0.0277/74.96	2
Sheep	52.9 km	1990/2010/2014	58,869	2.207/247.70	3
Goats	54.2 km	2000/2010/2014	51,039	0.695/418.30	4
Horses	54.1	1993/2010/2014	40,712	0.459/16.91	5
Pigs	43.3	2000/2010/2014	55,462	3.742/928.22	6
Chickens	43.2	1994/2010/2014	69,761	67.57/10,896	7
Ducks	54.4	1994/2009/2014	44,109	4.263/2,564	8

**Table 2.** List of data records, organised by species.

For the AW product, when pixels are smaller than the corresponding census polygons, all pixels are given an equal weight, and animals are distributed homogeneously within the census polygon, excluding unsuitable areas, where the animal density is set to 0. In pixels that are larger than the polygon size, the total animals from the polygons falling into the pixel are summed. In polygons where there were missing data in the global merge, the sum of the RF model mean prediction is used as an estimate of the total number of animals in the polygon, and these are distributed homogeneously in the same way as the observed census numbers.

In both the DA and AW products, pixels falling within polygons with missing animal totals are marked in a separate layer, so that users can distinguish densities derived from observed census numbers from those predicted by the RF models.

Finally, all pixels are corrected by a country factor so that the summed values of the pixel match the total number of animals registered in the FAOSTAT database for the reference year 2010. This ensures that subnational census data from different years are standardized to 2010 and that all totals are compatible with the numbers officially declared by countries to FAO. However, the original subnational country census data are provided in the metadata table. Users may revert to the original totals by applying the inverse country-level correction factor if needed.

### Code availability

The code is fully operational under R 3.3.3<sup>13</sup> and the key packages were raster 2.5–8<sup>14</sup>, rgdal 1.2–13<sup>15</sup>, maptools 0.9–2<sup>16</sup> and randomForest<sup>17</sup>. The full code used to implement GLW is available from the authors with no restriction but is currently provided with no detailed documentation.

### Data Records

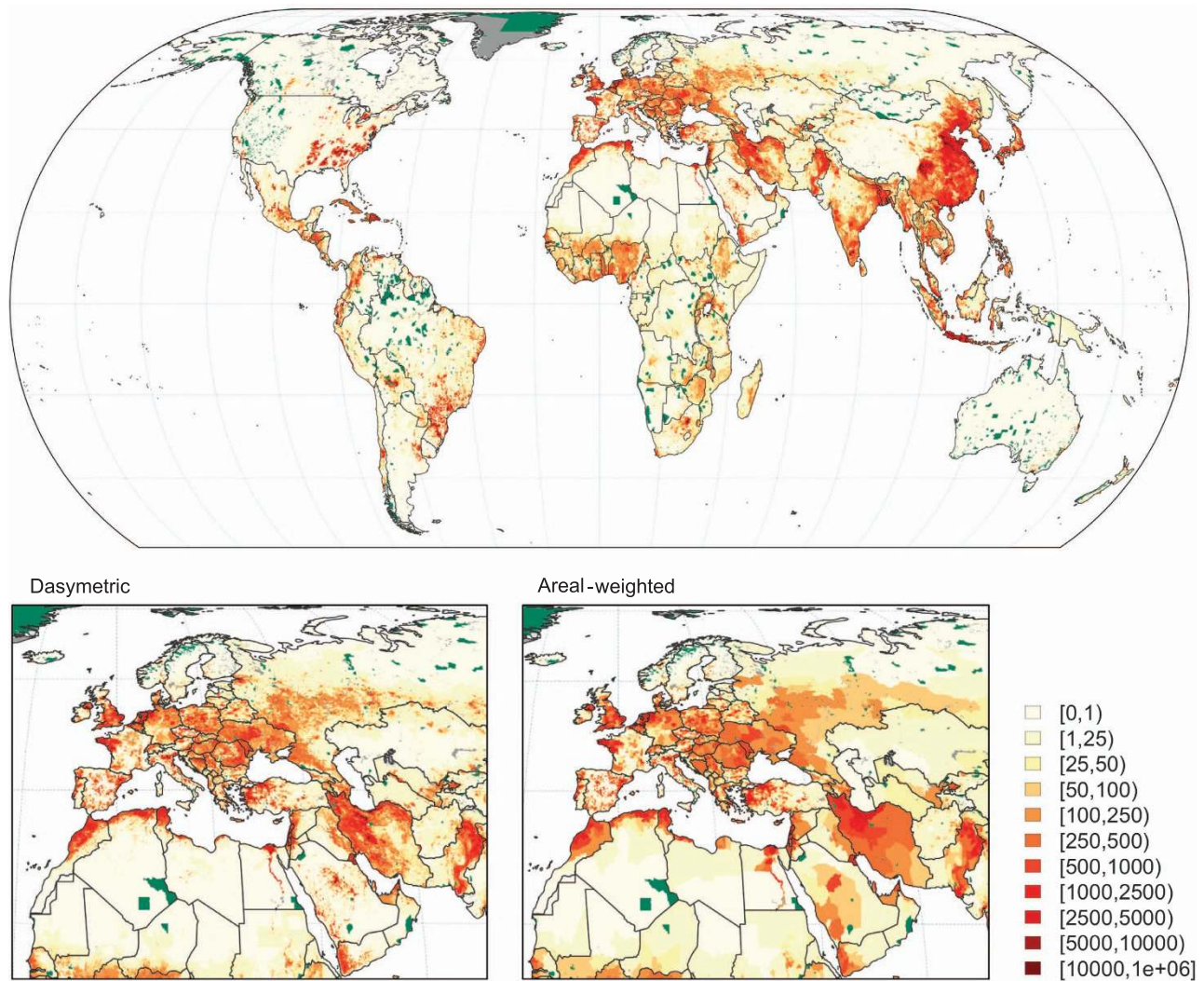
The data records described in this paper are publicly and freely available on the Gridded Livestock of World 3 Dataverse (Data Citation 1–8) and through the FAO livestock systems World Web site (<http://www.fao.org/livestock-systems/>). The data records are grouped by species (Table 2), and each species data record includes a metadata document, quick view graphic files and the GIS data as Geotiff files with a spatial extent of –180 to 180 degrees of longitude and –90 to 90 degrees of latitude. With a spatial resolution of 0.083333 decimal degrees per pixel, the resulting raster is 4,320 by 2,160 pixels (Table 3). The metadata document provides a detailed explanation of the different files, quick views of the different maps, ASR and census year maps and histograms, indicators of the RF models' GOF and a comprehensive list of original data sources grouped by countries and providing references to the publication and/or URL of the original country census data. Quick views and GIS raster files are provided for the dasymetric product, the areal-weighted product, and the distribution of prediction vs. observed status, highlighting areas where there were missing census data and where RF predictions were used (Table 3). As an example, Fig. 3 presents the predicted global distribution of chickens in the dasymetric product (top). The small inserts allow the difference between the DA (left) and AW (right) products to be observed. In countries where input census units are very small, such as Italy and Spain, the difference is hardly noticeable. In contrast, the AW product displays large areas with equal density in countries with large census units such as Russia and Iran, where the DA product redistributes chickens within census units according to the RF weights.

### Technical Validation

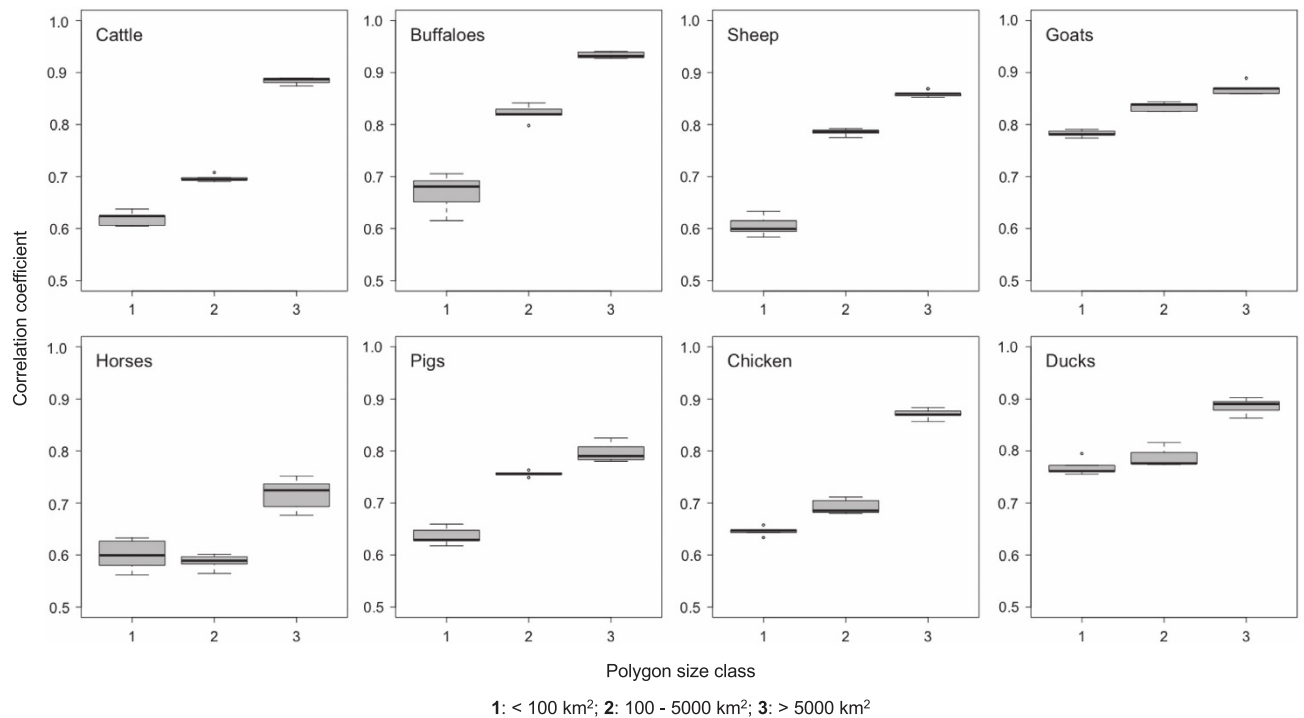
The technical validation was carried out internally by training the models with 70 percent of the input polygon data and evaluating the predictions using 30 percent of the polygons that were not used to train the model. The GOF was evaluated using both the RMSE and the correlation coefficient between the observed and predicted log-transformed numbers of animals per polygon. However, in the event that the RMSE and correlation coefficient were similar, we only report the correlation coefficient as an indicator of GOF. Figure 4 shows the GOF plot broken down by species and polygon size class. Individual GOF plots broken down by polygon size class and continent are provided in the individual metadata files for each species.

File name	Description	File type
1_Ch_2010_Metadata.html	Full metadata document report	Hypertext Markup Language
2_Ch_2010_Da.png	Quick view graphic file of the dasymetric product	Portable Network Graphics
3_Ch_2010_Aw.png	Quick view graphic file of the areal-weighted product	Portable Network Graphics
4_Ch_2010_Ps.png	Quick view graphic file of the prediction status of the areal-weighted product	Portable Network Graphics
5_Ch_2010_Da.tif	GIS file of the dasymetric product, (absolute number of animals per pixel; 4320 by 2160 pixels of 0.083333 decimal degrees resolution)	Geotiff
6_Ch_2010_Aw.tif	GIS file of the areal-weighted product, (absolute number of animals per pixel; 4320 by 2160 pixels of 0.083333 decimal degrees resolution)	Geotiff
7_Ch_2010_Ps.tif	GIS file of the areal-weighted prediction status (1: from predictions, 0: from observed census; 4320 by 2160 pixels of 0.083333 decimal degrees resolution)	Geotiff
8_Areackm.tif	GIS file of the area per pixel (square km; 4320 by 2160 pixels of 0.083333 decimal degrees resolution)	Geotiff

**Table 3.** List of files provided for each species, with chickens as an example. For other species, the “Ch” prefix is replaced by “Ct” for cattle, “Bf” for buffaloes, “Sh” for sheep, “Gt” for goats, “Ho” for horses, “Pg” for pigs and “Dk” for ducks.



**Figure 3.** Distribution of chicken density in the world and within Europe. The bottom panels highlight the difference between the dasymetric (bottom left) and areal-weighted (bottom right) databases. Dark grey are areas considered unsuitable and dark green areas correspond to IUCN protected areas.



**Figure 4.** Correlation coefficients between observed and predicted livestock densities broken down by polygon census size and species.

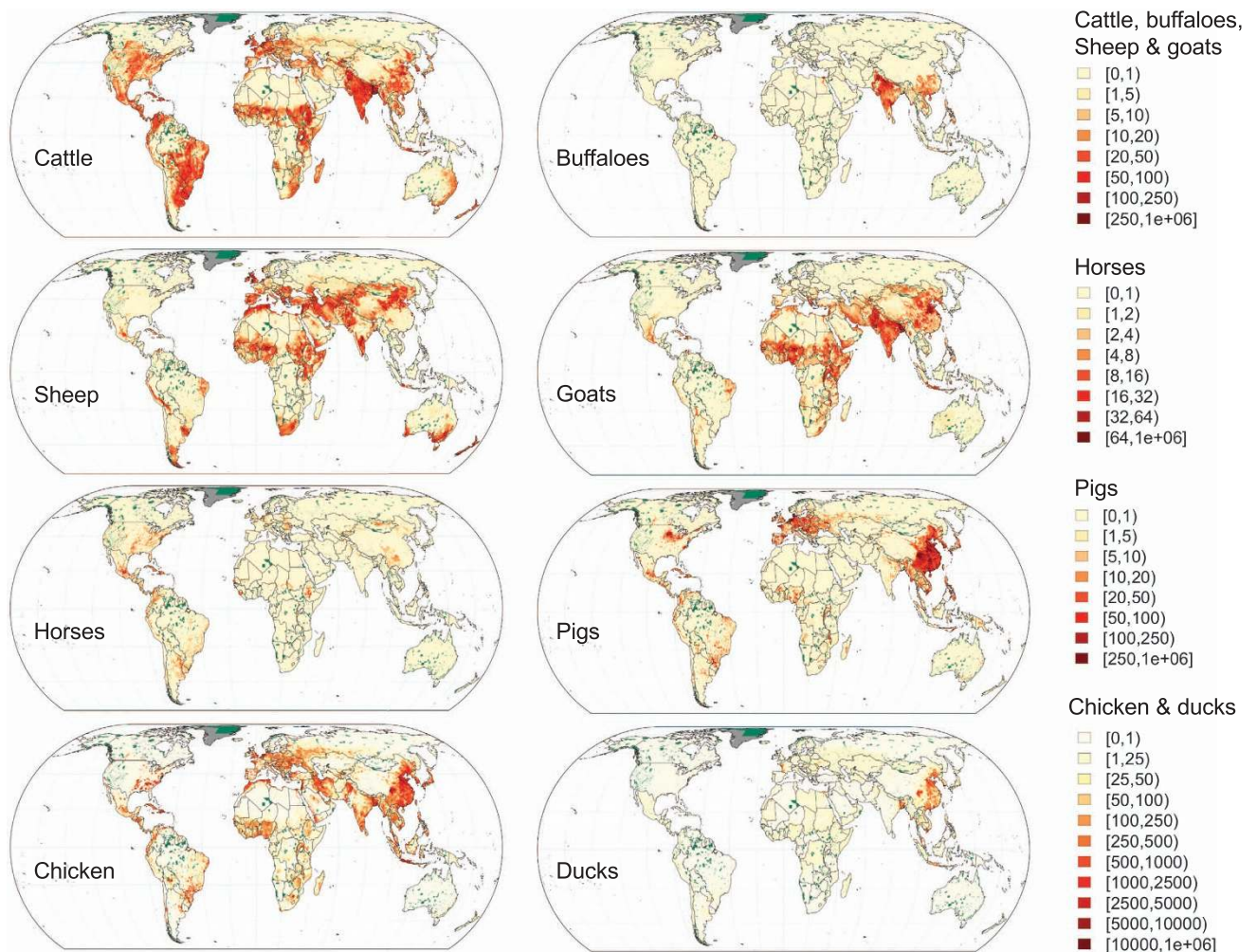
The GOF was moderate to high, depending on species and size of census administrative unit (Fig. 4). Since our spatial model predicts values at a spatial resolution of roughly 10 km (at the equator), the GOF metric of the first polygon size class ( $< 100\text{km}^2$ ) gives the most accurate estimate of the prediction accuracy of the pixel-level predictions. These ranged between 0.60 and 0.78, meaning that a significant part of the variability is not captured by the model and that pixel-level estimates cannot be assumed to fully represent what is on the ground. This could be linked to important predictor variables that are absent from the model, or to the stochastic nature of the spatial allocation of farms. As spatial units become larger, the variability gets filtered out and the observed values become easier to predict, with correlation coefficients close to 0.90 for the largest units. So, the gap-filling capacity of the models can be assumed to be good, which benefits both the DA and AW products. The GOF metrics by continent and species sometime reflect very different model qualities depending on the continent. Future studies should evaluate whether alternative stratifications could better harmonize the quality of the models across geographic regions. For example, groupings based on the economic status of countries may prove more appropriate than those based on continents. Livestock farming is constrained by production factors such as land, capital and manpower, and the last two are strongly associated with countries' economic status, which influences how farms can be distributed across the landscape.

The GOF metrics need to be interpreted with care because they result from internal cross-validation. They do not measure the correspondence between the predicted densities of animals and what is actually on the ground. If the census itself is of poor quality, there could be discrepancies between the recorded numbers and what is actually on the ground, let alone between the census and the predicted values. Furthermore, census data are mostly based on where animals are registered to their owner, not necessarily where they are raised or spend most of their lifetime. For ruminants raised in pastoral systems, or ducks raised in free-grazing systems, for example, there could be significant seasonal changes in the spatial distributions of animals that would not be captured in the models.

The outputs also assume no livestock to occur in IUCN protected areas, and imposing a density of zero in such areas. It is, however, known that livestock encroach on protected areas and the validity of these assumptions depends on how effectively these restrictions are enforced. This varies greatly from country to country and even within countries.

One possible way to validate the models would be to use household demographic and health surveys (DHS) or living standards measurement studies (LSMS) data on livestock ownership. These data are typically geo-referenced at the cluster level, follow a completely different sampling approach and would therefore provide a different base for the development of livestock models. Cross-checking the results of models derived from large-scale census and from point-based surveys would help to identify areas of convergence where predictions would be consolidated, and areas of divergence where there would be





**Figure 5.** Overview of the Gridded Livestock of the World (GLW 3) data sets for cattle, buffaloes, sheep, goats, horses, pigs, chickens and ducks, based on the dasymetric model. Dark grey are areas considered unsuitable and dark green areas correspond to IUCN protected areas.

higher uncertainties associated with the predictions. Field observations and aerial surveys provide efficient means of collecting high-quality data that could be used to validate the models, however, they are costly, especially when carried out over large areas. The cost-effectiveness of remote sensing for counting large animals was recently reviewed, and appears currently to be of limited utility<sup>18</sup>.

### Usage Notes

This version of GLW is suited to applications in the domains of socio-economics, environment and health. The data are most appropriate for applications at global and continental scales. Decisions regarding the use of this version of GLW over smaller spatial extents should be taken in relation to the ASR of the underlying census data. For example, analyses of GLW data in Brazil, Spain or Thailand would be appropriate because their respective ASRs are small relative to the size of the country. In contrast, we would discourage the country-level use of GLW data in countries such as Russia or Mali, because the ASR values are particularly high (>250 km). It is also important that users of these data are mindful of the fact that the type of production system is not accounted for in these livestock distribution data. The diverse contexts in which livestock are raised have major bearings on their primary uses, their productivity, the benefits they confer, the constraints to production and the impacts they have. The distribution data should therefore be used in conjunction with information on production systems. Currently, global<sup>1</sup> and regional<sup>19</sup> ruminant production systems data are available and global monogastric systems are available for pigs and chickens<sup>20</sup>.

The DA version is recommended for applications where spatial detail matters more than concerns about circularity in the analytical workflow in relation to spatial predictor covariates. However, we warn potential users against over-interpretation of spatial accuracy of the DA product. As indicated by the GOF metrics, much variability was not captured by the models and the downscaled densities of the DA

version only imperfectly represent what may actually be on the ground. When circularity concerns are more important than detailed spatial resolution, it is recommended that the AW versions be used.

This is the third version of GLW, and the previous version had reference years of 2002 and 2006, respectively. However, since the three versions of GLW differ in the type of input data, the predictor covariates and modelling methods, we would discourage their use for time-series analysis. Future studies will develop appropriate models to map how livestock distributions have changed over time.

In order to facilitate zonal summations, the values in each pixel of the DA and AW data sets correspond to the absolute numbers of animals, not to densities. These values can be converted to densities (number per km<sup>2</sup>) by dividing each pixel value by the pixel area in km<sup>2</sup>. For convenience, a global Geotiff file of pixel areas, expressed in km<sup>2</sup>/pixel, is provided along with each species data file (Table 3).

All outputs have been corrected so that the total number of animals in a country matches the FAOSTAT 2010 total stock. However, in a number of cases, there are significant differences between the total numbers of animals found in the original national census data and the values recorded in FAOSTAT. The total of the original census is provided for each country and species in the metadata report so that users may revert to these by dividing all pixel values by the FAOSTAT 2010/Census total ratio. Figure 5 shows the global distributions of the eight livestock species included based on the DA models.

## References

1. Robinson, T. *et al.* Global livestock production systems. 152 pp (2011).
2. Steinfeld, H., Gerber, P., Wassenaar, T. D., Castel, V. & de Haan, C. *Livestock's long shadow: environmental issues and options.* (FAO, 2006).
3. Slingenbergh, J. & others. *World Livestock 2013: changing disease landscapes.* (Food and Agriculture Organization of the United Nations FAO, 2013).
4. Gilbert, M., Xiao, X. & Robinson, T. P. Intensifying poultry production systems and the emergence of avian influenza in China: a 'One Health/Ecohealth' epitome. *Arch. Public Health* **75**, 48 (2017).
5. Van Boeckel, T. P. *et al.* Global trends in antimicrobial use in food animals. *Proc. Natl. Acad. Sci* **112**, 5649–5654 (2015).
6. Van Boeckel, T. P. *et al.* Reducing antimicrobial use in food animals. *Science* **357**, 1350–1352 (2017).
7. Wint, W. & Robinson, T. *Gridded Livestock of the World.* (Food and Agriculture Organization, 2007).
8. Robinson, T. P. *et al.* Mapping the Global Distribution of Livestock. *PLoS ONE* **9**, e96084 (2014).
9. Nicolas, G. *et al.* Using Random Forest to Improve the Downscaling of Global Livestock Census Data. *PLOS ONE* **11**, e0150424 (2016).
10. Channan, S., Collins, K. & Emanuel, W. R. Global mosaics of the standard MODIS land cover type data. (University of Maryland and the Pacific Northwest National Laboratory, 2014).
11. Pesaresi, M. *et al.* *GHS Built-up Grid, Derived from Landsat, Multitemporal (1975, 1990, 2000, 2014)* European Commission, Joint Research Centre (JRC). PID. (2015).
12. Dudley, N. *Guidelines for applying protected area management categories.* (Iucn, 2008).
13. R. Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing: Vienna, Austria, 2012. (ISBN 3-900051-07-0, 2012).
14. Hijmans, R. J. & van Etten, J. raster: Geographic data analysis and modeling. *R Package Version 2*, 15 (2014).
15. Bivand, R. *et al.* Package 'rgdal'. (2017).
16. Bivand, R. *et al.* Package 'mapproj'. (2017).
17. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
18. Hollings, T. *et al.* How do you find the Green Sheep? A critical review of the use of remotely sensed imagery to detect and count animals. *Methods Ecol. Evol* **9**, 881–892 (2018).
19. Cecchi, G. *et al.* Geographic distribution and environmental characterization of livestock production systems in Eastern Africa. *Agric. Ecosyst. Environ* **135**, 98–110 (2010).
20. Gilbert, M. *et al.* Income Disparities and the Global Distribution of Intensively Farmed Chicken and Pigs. *PLoS ONE* **10**, e0133381 (2015).
21. Center for International Earth Science Information Network - CIESIN - Columbia University. *Gridded Population of the World, Version 4 (GPWv4): Land and Water Area.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), (2016).
22. UNEP-WCMC and IUCN. *The World Database on Protected Areas (WDPA).* (UNEP-WCMC, 2010).
23. Tatem, A. J. WorldPop, open data for spatial demography. *Sci. Data* **4**, 170004 (2017).
24. Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C. & Worley, B. A. LandScan: a global population database for estimating populations at risk. *Photogramm. Eng. Remote Sens.* **66**, 849–857 (2000).
25. Center for International Earth Science Information Network - CIESIN - Columbia University. *Gridded Population of the World, Version 4 (GPWv4): Population Count.* Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), (2016).
26. Nelson, A. *Travel time to major cities: A global map of Accessibility.* Global Environment Monitoring Unit—Joint Research Centre of the European Commission: Ispra, Italy. (2008).
27. LDAAC. *Global 30 Arc-Second Elevation Data Set GTOPO30.* (Land Process Distributed Active Archive Center, 2004).
28. Scharlemann, J. P. *et al.* Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PLoS One* **3**, e1408 (2008).
29. Jones, P. G. & Thornton, P. K. Croppers to livestock keepers: livelihood transitions to 2050 in Africa due to climate change. *Environ. Sci. Policy* **12**, 427–437 (2009).
30. Zhang, X. *et al.* Monitoring vegetation phenology using MODIS. *Remote Sens. Environ.* **84**, 471–475 (2003).
31. Fritz, S. *et al.* Mapping global cropland and field size. *Glob. Change Biol.* **21**, 1980–1992 (2015).
32. Hansen, M. C. *et al.* High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* **342**, 850–853 (2013).
33. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).

## Data Citations

1. Gilbert, M. *et al.* *Harvard Dataverse* <http://dx.doi.org/10.7910/DVN/GIVQ75> (2018).
2. Gilbert, M. *et al.* *Harvard Dataverse* <http://dx.doi.org/10.7910/DVN/5U8MWI> (2018).
3. Gilbert, M. *et al.* *Harvard Dataverse* <http://dx.doi.org/10.7910/DVN/BLWPZN> (2018).

4. Gilbert, M. *et al.* *Harvard Dataverse* <http://dx.doi.org/10.7910/DVN/OCPH42> (2018).
5. Gilbert, M. *et al.* *Harvard Dataverse* <http://dx.doi.org/10.7910/DVN/7Q52MV> (2018).
6. Gilbert, M. *et al.* *Harvard Dataverse* <http://dx.doi.org/10.7910/DVN/33N0JG> (2018).
7. Gilbert, M. *et al.* *Harvard Dataverse* <http://dx.doi.org/10.7910/DVN/SUFASB> (2018).
8. Gilbert, M. *et al.* *Harvard Dataverse* <http://dx.doi.org/10.7910/DVN/ICHCBH> (2018).

### Acknowledgements

This work is supported by the FAO Animal Production and Health Division's regular programme and by a number of projects within that. Significant contributions to this project have been made by the Belgian FNRS projects "Mapping Livestock and People" PDR T.0073.13 and "Mapping Livestock transition" WISD X.3023.17. The views expressed here are those of the authors and not necessarily those of FAO or any other organization.

### Author Contributions

M.G. and T.R. designed the study. M.G., G.N. & G.C. developed and implemented the RF modelling of the different data sets. G.C. & W.W. provided and managed key input census and covariate data. All authors contributed to the final paper.

### Additional Information

**Competing interests:** The authors declare no competing interests.

**How to cite this article:** Gilbert, M. *et al.* Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci. Data*. 5:180227 doi: 10.1038/sdata.2018.227 (2018).

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files made available in this article.

© The Author(s) 2018