SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

CassavaMap, a fine-resolution disaggregation of cassava production and harvested area in Africa in 2014

Anna Maria Szyniszewska ⓘD

Cassava, the third main source of carbohydrates in Africa, provides daily nutrition for over 700 million people on the continent and a vital source of income for subsistence farmers. Despite its importance, our knowledge of the heterogeneity of its distribution in the landscape is limited and outdated. Information on cassava production and harvested area are typically available on aggregated administrative unit level with highly variable temporal range of records, often over a decade old, and represented on a coarse ~10 km by 10 km grid. Here, cassava production and harvested area administrative unit level data for 32 countries are standardised to 2014 FAO reported levels and disaggregated based on the distribution of the rural population in 2014. The grid obtained represents a significant improvement on the previous studies in terms of both spatial resolution (~1 km by 1 km) and temporal accuracy. Enhanced representation of cassava production and harvested area in Africa is an essential resource for policy making as well as designing strategies to manage its main pathogens.

Background & Summary

Cassava (*Manihot esculenta* Krantz) is a root crop grown in the tropics. It is a rich source of calories and provides several important micronutrients[1,2]. Cassava is grown mainly for subsistence and direct consumption, but has a growing market of processing opportunities[3]. Currently, its production in sub-Saharan Africa is drastically hindered by several pests and diseases. The most devastating of these are the virus diseases – cassava mosaic disease (CMD), and cassava brown streak disease (CBSD)[4–7].

Considerable resources have been invested by various agricultural development donors in surveillance activities for those diseases, as well as in the development and deployment of control strategies, including clean seed and resistant varieties[8–10]. In all of these activities, a reliable representation of the distribution of cassava cultivation is an essential resource for planning. However, our knowledge about the spatial distribution of cassava from currently existing published studies or output from governments of international agencies such as FAO is outdated, and often represented only by national or sub-national level statistics for cassava production or harvested area[11].

There are two resources known to the author that represent the results of the modelling of the distribution of cassava production and harvested area based on agriculture census data and disaggregating it to grid. The first model, available via the CGIAR Research Program on Roots, Tubers and Bananas (RTB)[12], provides information on the spatial distribution of numerous crops including cassava. More specifically, it provides information on the spatial distribution of cassava harvested area, yield gap, potential yield and so called cassava suitability index with a spatial resolution at the equator of approximately 10 km by 10 km. These maps were obtained by disaggregating statistics on harvested areas and yields of 175 crops among cropland areas represented by two different land cover classification sources: MODIS and GLC2000, and represent production statistics for ca. 2000. The second source of information on the distribution of crops is the Spatial Production Allocation Model (MapSPAM)[13–15]. MapSPAM uses a variety of inputs (crop production statistics, irrigated land, croplands, rural population and crop biophysical suitability index) to disaggregate the production of 42 crops including cassava simultaneously using an entropy-based data-fusion approach[14]. At the time of writing this paper, the newest version of MapSPAM was representing crop production statistics for 2000, 2005 and 2010, depending on the version, at ~10 km by

Department of Plant Sciences, University of Cambridge, Cambridge, UK. e-mail: aniasz@gmail.com
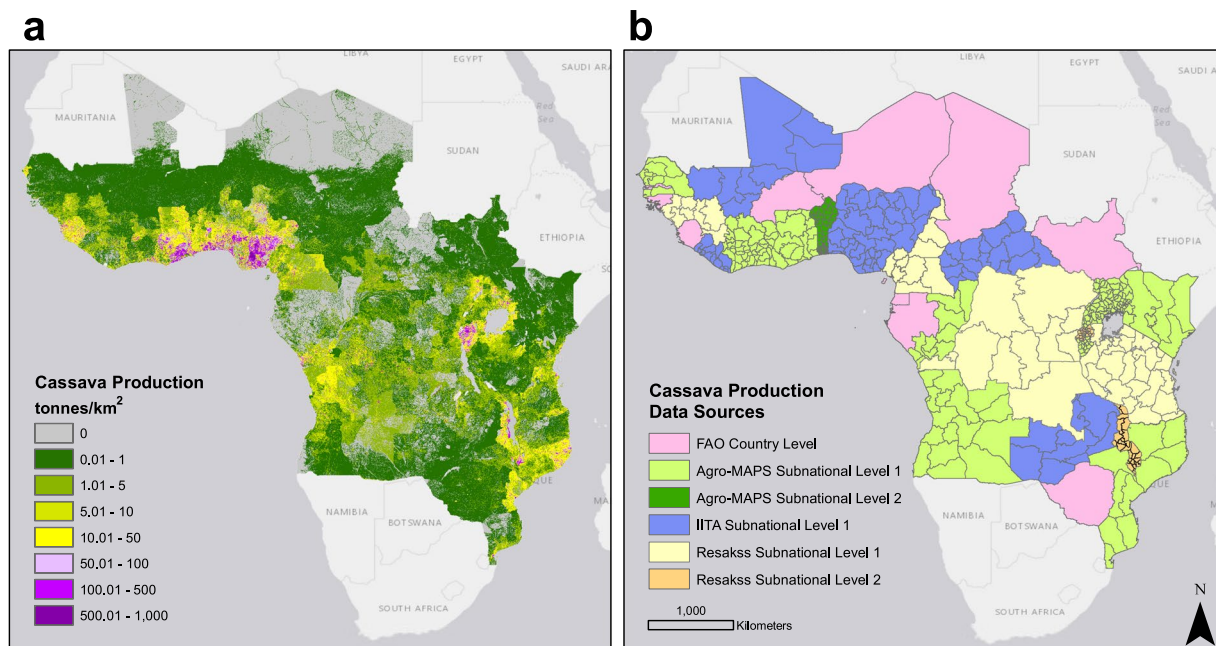
Fig. 1 Disaggregated cassava production in sub-Saharan Africa and associated data sources. Density of cassava production layer (tonnes per approx. 1 km²) (**a**). Cassava production data sources and administrative resolution of production census data in each of the countries (**b**): FAO (fao.org/faostat), Resakss (resakss.org), Agro-MAPS (kids.fao.org/agromaps/) and IITA (personal correspondence with the International Institute of Tropical Agriculture in Ibadan, Nigeria). Basemap source: Esri, DeLorme, HERE, MapmyIndia.

10 km resolution globally or ~1 km by 1 km resolution on a country level for some countries. This model used data on harvested area to estimate the areas of cropland in each administrative unit.

Mapping cassava on a continental scale based on satellite imagery remains a major challenge. Cassava in sub-Saharan Africa dominates in small-holder farming systems with field sizes most frequently less than 1000 m², which are often intercropped[16,17]. Field rotation, flexible harvesting times and variable planting seasons means there is no strongly defined pattern of field maturation in many areas. In addition, the variability of cultivars and (most importantly) agro-ecological zones means the cassava spectral signal and texture pattern on satellite imageries varies greatly. The combination of these challenges motivated this work. The objective was to find an alternative approach of estimating cassava production and harvested area density based on available proxy variables.

Previous studies have shown that the distribution of non-urban population is an important predictor of the density of cassava in the landscape in sub-Saharan Africa, and it's relative importance varies depending on regional cultural preferences[18–21]. The crop is adaptable to various environmental conditions and there are no limiting factors in terms of its distribution like temperature or rainfall in the majority of the study region. Unlike rice and maize, it is highly drought resistant and can be grown on marginal soils[18]. Due to the limited number of non-environmental ancillary covariates that are likely to influence the preference of cassava planting in sub-Saharan Africa like cultural preferences or poverty, the distribution of rural population was assumed to be the main driver of cassava density in the landscape and the relationship between population number and cassava production and harvested area density linear up to defined maximum threshold. Available cassava production and harvested area data obtained from various existing sources including FAO, ReSAKSS, and International Institute of Tropical Agriculture (IITA) communication were used as proxies representing regional importance of the crop. The administrative level cassava production and harvested area data were disaggregated into small spatial units (pixels) at approximately 30″ × 30″ (~1 × 1 km) resolution, between rural population data obtained from LandScan 2014. The two layers obtained provide the finest up-to-date representation of cassava distribution in the landscape in 30 countries of sub-Saharan Africa currently available (Figs. 1 and 2).

Methods

Rural population.   Representation of human population density was used to redistribute data on cassava production per province or district in 30 countries of West, central, East and southern Africa. Data were obtained from the LandScan 2014[22] raster population density layer developed as part of the Oak Ridge National Laboratory (ORNL) Global Population Project at a resolution of approximately at ~1 km by 1 km (~30″ by 30″). The LandScan model uses country sub-national census data and combines it with additional datasets including land cover, roads, urban and rural locations to redistribute the populations according to a weighting scheme. Populations living in densely populated urban areas are unlikely to grow cassava and thus, areas with a population density higher than 5,000 inhabitants per 1 km² were masked out and deemed unsuitable for cassava farming.
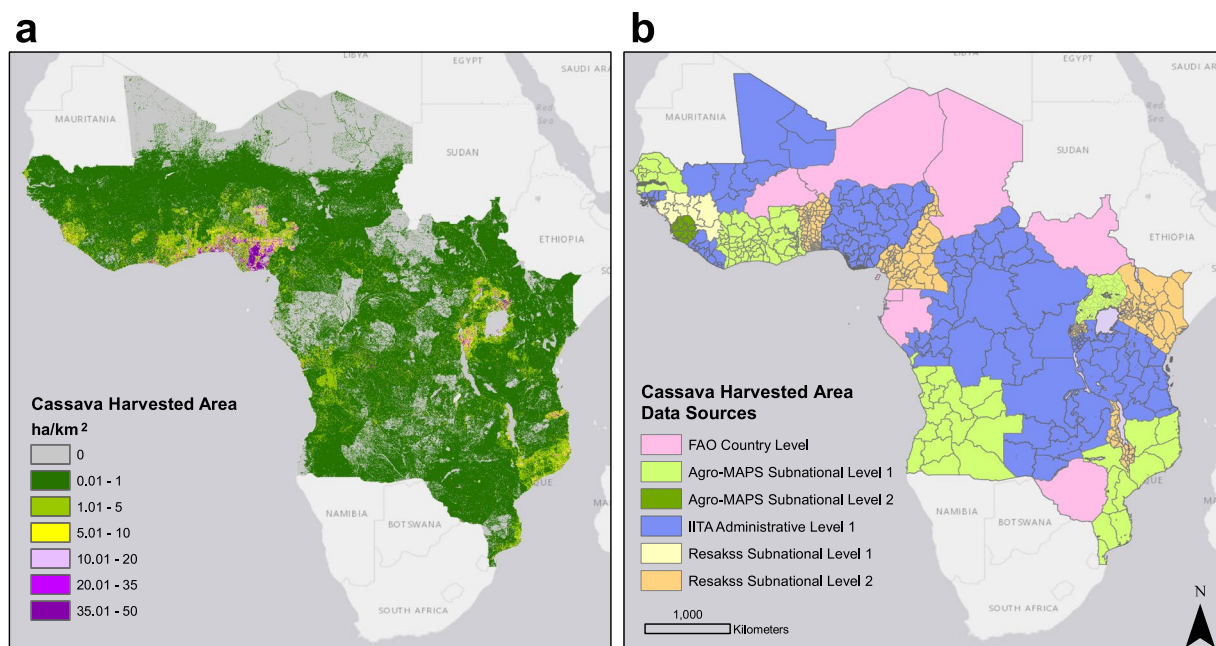
Fig. 2 Disaggregated cassava harvested area in sub-Saharan Africa and associated data sources. Density of cassava harvested area layer (hectares per approx. 1 km²) (**a**) and harvested area data sources as well as administrative resolution of data in each of the countries (**b**): FAO (fao.org/faostat), Resakss (resakss.org), Agro-MAPS (kids.fao.org/agromaps/) and IITA (personal correspondence with the International Institute of Tropical Agriculture in Ibadan, Nigeria). Basemap source: Esri, DeLorme, HERE, MapmyIndia.

Cassava production and harvested area data.   Sub-national data about cassava production and harvested area per administrative unit were obtained and compiled from several sources: FAO Agro-MAPS (www.kids.fao.org/agromaps), ReSAKSS (www.resakss.org) and personal communication with the International Institute of Tropical Agriculture (IITA). These sources were later compared according to the administrative resolution, year of data records and amount of missing data. Based on these three criteria, the preferred data source was manually identified to strike the best balance most recent records, fine resolution and finally considering number of missing records (Online-only Tables 1 and 2). For some countries no sub-national statistics on cassava production or harvested area were available. In those instances, country level FAO estimates (www.fao.org/faostat) for 2014 were used for cassava production and harvested area estimates (Online-only Tables 3 and 4). Collated data represented statistics from various years, as agricultural censuses are taken rarely, often every 10 years or even less frequently. Country-level FAO cassava production and harvested area estimates from 2014 were used to standardise values in the administrative units across each country so the totals represented those of the 2014 national-level estimates.

Administrative units.   Due to frequent changes in administrative divisions in various African countries, polygon shapefiles provided with respective sources of cassava production statistics were used to represent the extent of administrative zones. Administrative division shapefiles for Uganda provided by FAO Agro-MAPS were unfit for use due to the corrupted topology of the layer and the lack of a match between the unique identifiers and names of the districts listed in the production data file. In this case, the country administrative shapefile for the first level of administrative division was downloaded from GADM (www.gadm.org). Unions between districts that have been historically merged were applied in order to link them correctly to the Agro-MAPS statistics file.

Missing data.   In multiple subnational administrative units, production, or harvested area data respectively were missing (Online-only Tables 3 and 4). Based upon personal communication with country experts it was established that cassava is grown in those areas. In those areas cassava per capita production/harvested area values were calculated individually based either on the average value of all administrative units from the same country adjacent to the unit with missing data or based on values in districts representing the same or similar agro-ecological zones (Online-only Table 5).

The Ugandan country outline shapefile obtained from GADM had relatively small areas that lacked alignment with neighbouring Kenya and Democratic Republic of Congo (DRC) shapefiles. It resulted in two areas containing a proportion of pixels with missing values within the bounding boxes defined by the following sets of coordinates: 35.797°E, 4.146°N, 33.963°W, 1.569°S and (30.993°E, 3.463°N, 30.742°W, 2.514°S. Missing value pixels in these bounding boxes were replaced with a constant value obtained from the range of values in neighbouring pixels (0.02 for production and harvested area per capita 2014 adjusted values).

Disaggregation.   For each administrative unit, cassava production in tonnes and harvested area in hectares per capita values adjusted for 2014 FAO reported country totals (*Prod_percapita_2014* and *HA_percapita_2014*) were calculated. They were obtained by dividing the adjusted production total (*Admin_unit_prod2014* and *Admin_unit_ha2014*) (Eqs. 1 and 2) by the rural population (*Admin_unit_pop2014*) (areas at or below 5,000 inhabitants per pixel) in each administrative unit (Eqs. 3 and 4).

$$Admin\_unit\_prod\_2014 = \frac{Admin\_unit\_prod}{Country\_prod} \times Country\_prod2014 \tag{1}$$

$$Admin\_unit\_ha\_2014 = \frac{Admin\_unit\_ha}{Country\_ha} \times Country\_ha2014 \tag{2}$$

$$Prod\_percapita\_2014 = \frac{Admin\_unit\_prod\_2014}{Admin\_unit\_pop2014} \tag{3}$$

$$HA\_percapita\_2014 = \frac{Admin\_unit\_ha\_2014}{Admin\_unit\_pop2014} \tag{4}$$

A linear relationship between population number and cassava production was assumed up to a maximum allowed threshold value. The per capita cassava production and harvested area values were evenly allocated across all pixels by multiplying the number of people by the per capita value. It was assumed that a maximum of half of the area of the pixel can be allocated to cassava production (approx. 50 hectares) and the production per hectare does not exceed 20 tonnes per hectare. Thus, cassava harvested area per pixel cannot exceed 50 hectares and cassava production per pixel cannot exceed 1,000 tonnes. For those districts where cassava harvested area exceeded 50 hectares or production exceeded 1,000 tonnes in one or more pixels, the amount in excess was deducted and redistributed between the remaining pixels until the condition was satisfied and no pixel in a given administrative unit exceeded the assumed maximum production/harvested area threshold value.

Data Records
CassavaMap provides representation of cassava production and harvested area in 32 countries of sub-Saharan Africa (Figs. 1 and 2). CassavaMap version 1.0 can be downloaded from the Figshare repository under CC BY 4.0 licence[23]. Data records constitute two raster TIFF files representing disaggregated cassava production (CassavaMap_Production_v1.tif) and harvested area (CassavaMap_HarvArea_v1.tif). CassavaMap_Prod_v1.csv and CassavaMap_HarvArea_v1.csv are data table files with an outline of detailed production and harvested area statistics in 474 and 644 administrative units respectively. Both files include information on source of cassava statistics data, administrative level, year of data records, rural population sum per administrative unit derived from LandScan 2014, summarised country production/harvested area, FAOSTAT production/harvested area per country in 2014, per capita production/harvested area, and per capita production/harvested area values adjusted to national production levels in 2014 as listed by FAOSTAT. CassavaMap_Legend.csv file explains the column headers for both data table files. Missing_data_input_calc.csv outlines the administrative units with missing data replaced by estimated values. The unique names or codes of neighbouring administrative units that were used to derive the new input per capita production or harvested area values are outlined in respective rows and the function used to estimate those values is listed. Newer versions of CassavaMap will appear in the repository as updated cassava production data become available.

Technical Validation
CassavaMap represents a smooth fine-resolution disaggregation of cassava production and harvested area based on assumption that the distribution of rural population is the main driver of cassava density in sub-Saharan Africa. It is important to mention that model's assumptions and its scale make it prone to under- or over-representation of cassava production if the underlying data are not accurate. First source of errors may be inherited from the LandScan 2014 population distribution model, which is based on a set of algorithms for redistributing population in given census zones. Second source of misrepresentation may occur where production statistics as reported by sources used in this study are over- or under-estimated. Finally, areas with coarse administrative granularity of data (especially these where data were available only on national scale), will have production/harvested area redistributed in a flatter manner underrepresenting regional variations in cassava planting intensity. Several areas of sharp contrast in cassava production or harvested area levels include Zambia/Malawi and Ghana/Burkina Faso borders.

CassavaMap v1.0 improves 10-fold the spatial resolution of the currently available crop distribution models and standardises the output to 2014 country-wide production levels reported by FAO.

Code availability
The data analysis was performed in R version 3.5.1[24]. The example of the custom code used to disaggregate cassava production and harvested area data is publically available via GitHub at https://github.com/aniaszy/CassavaMap.

References

1. Chavez, A. L. *et al.* Iron, Carotene, and Ascorbic Acid in Cassava Roots and Leaves. *Food Nutr. Bull.* **21**, 410–413 (2000).
2. Maziya-Dixon, B., Kling, J. G., Menkir, A. & Dixon, A. Genetic Variation in Total Carotene, Iron, and Zinc Contents of Maize and Cassava Genotypes. *Food Nutr. Bull.* **21**, 419–422 (2000).
3. Kolawole, P. O., Agbetoye, L. & Ogunlowo, S. A. Sustaining World Food Security with Improved Cassava Processing Technology: The Nigeria Experience. *Sustainability* **2**, 3681–3694 (2010).
4. Samura, A. E., Lakoh, K. A., Nabay, O., Fomba, S. N. & Koroma, J. P. Effect of Cassava Mosaic Disease (CMD) on Yield and Profitability of Cassava and Gari Production Enterprises in Sierra Leone. *J. Agric. Sci* **9**, 205 (2017).
5. Patil, B. L., Legg, J. P., Kanju, E. & Fauquet, C. M. Cassava brown streak disease: a threat to food security in Africa. *J. Gen. Virol* **96**, 956–968 (2015).
6. Legg, J. P. *et al.* Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa. *Virus Res.* **159**, 161–170 (2011).
7. Legg, J. P., Owor, B., Sseruwagi, P. & Ndunguru, J. Cassava Mosaic Virus Disease in East and Central Africa: Epidemiology and Management of A Regional Pandemic. *Adv. Virus Res* **67**, 355–418 (2006).
8. Legg, J. P. *et al.* Community phytosanitation to manage cassava brown streak disease. *Virus Res.* **241**, 236–253 (2017).
9. McQuaid, C. F. *et al.* Spatial dynamics and control of a crop pathogen with mixed-mode transmission. *PLOS Comput. Biol.* **13**, e1005654 (2017).
10. McQuaid, C. F., Sseruwagi, P., Pariyo, A. & Bosch, F. van den. Cassava brown streak disease and the sustainability of a clean seed system. *Plant Pathol.* **65**, 299–309 (2018).
11. FAOSTAT. Statistical data. *Food and Agriculture Organization of the United Nations, Rome* www.fao.org/faostat (2018).
12. CIAT, CIT, Biodiversity International & IITA. RTB Maps, http://www.rtb.cgiar.org/RTBMaps/ (2014).
13. International Food Policy Research Institute. Global Spatially-Disaggregated Crop Production Statistics Data for 2000 Version 3.0.7. *Harvard Dataverse* https://doi.org/10.7910/DVN/A50I2T (2019).
14. You, L. *et al.* Spatial production allocation model (SPAM) 2005 v2.0, http://mapspam.info (2014).
15. International Food Policy Research Institute. Global Spatially-Disaggregated Crop Production Statistics Data for 2005 Version 3.2. *Harvard Dataverse* https://doi.org/10.7910/DVN/DHXBJX (2019).
16. Night, G. *et al.* Occurrence and distribution of cassava pests and diseases in Rwanda. *Agric. Ecosyst. Environ* **140**, 492–497 (2011).
17. Szyniszewska, A. M. *et al.* Cassava planting material movement and grower behaviour in Zambia: implications for disease management. Preprint at, https://www.biorxiv.org/content/10.1101/528851v2 (2019).
18. Carter, S. E. & Jones, P. G. A model of the distribution of cassava in Africa. *Appl. Geogr.* **13**, 353–371 (1993).
19. Campo, B. V. H., Hyman, G. & Bellotti, A. Threats to cassava production: known and potential geographic distribution of four key biotic constraints. *Food Secur* **3**, 329–345 (2011).
20. Ugwu, B. O. & Nweke, F. I. Determinants of cassava distribution in Nigeria. *Agric. Ecosyst. Environ* **60**, 139–156 (1996).
21. Carter, S. E. *An atlas of cassava in Africa: historical, agroecological and demographic aspects of crop distribution.* (CIAT, 1992).
22. Bright, E. A., Rose, A. N. & Urban, M. L. *LandScan 2015 High-Resolution Global Population Data Set*, http://www.osti.gov/scitech/biblio/1340997 (2016).
23. Szyniszewska, A. M. CassavaMap. *figshare* https://doi.org/10.6084/m9.figshare.9745118 (2020).
24. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, https://www.r-project.org/ (2018).

Acknowledgements

Author contributions

A.M.S. conceived the idea, analysed data and drafted the manuscript.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.M.S.

Reprints and permissions information is available at www.nature.com/reprints.

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