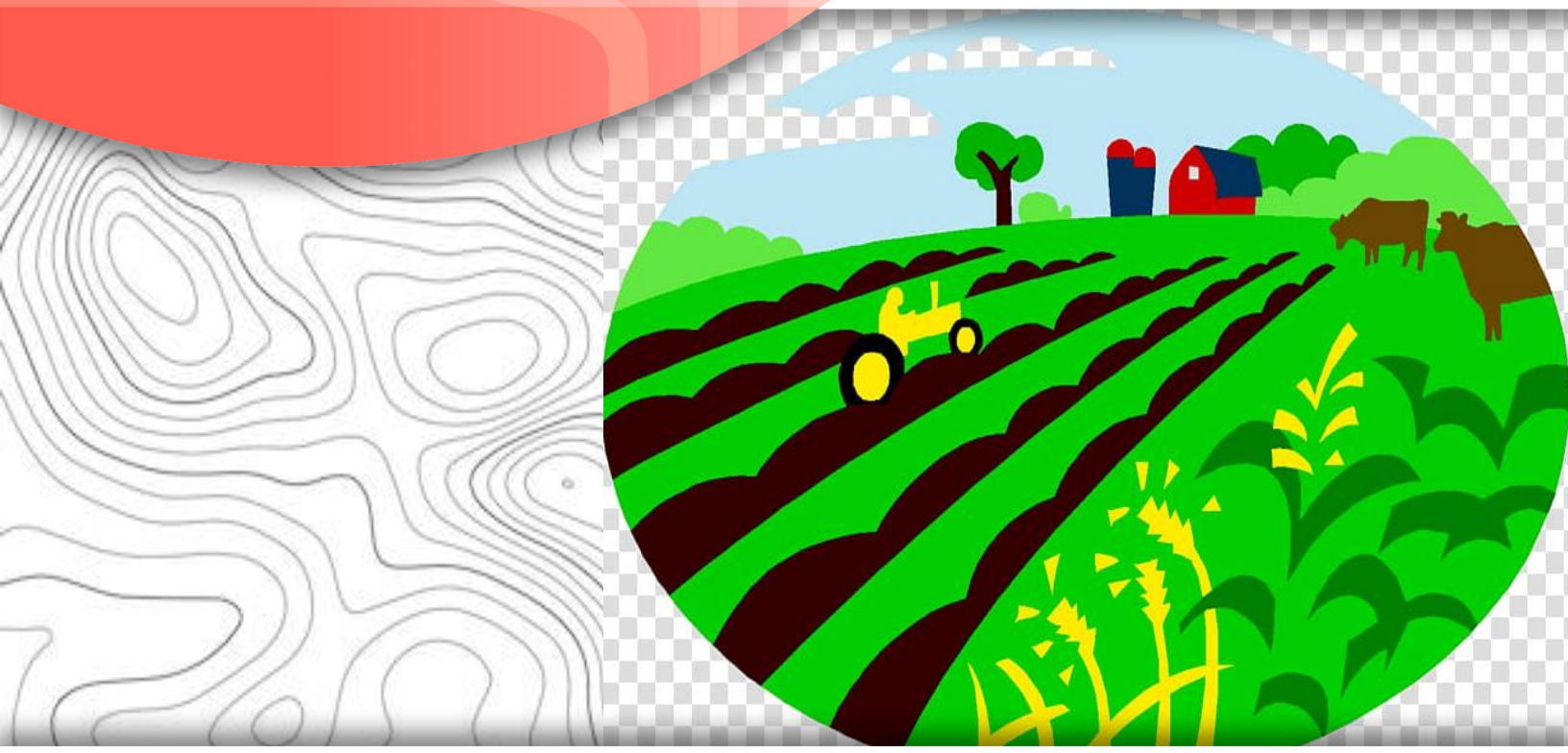
CGIAR

INITIATIVE ON

Market Intelligence



2023

Acknowledgment
This report was produced as part of the CGIAR Initiative on Market Intelligence and is supported by contributors to the CGIAR Trust Fund (https://www.cgiar.org/funders).

About CGIAR
CGIAR is a global research partnership for a food-secure future. CGIAR science is dedicated to transforming food, land and water systems in a climate crisis. Its research is carried out by 13 CGIAR Centers/Alliances in close collaboration with hundreds of partners, including national and regional research institutes, civil society organizations, academia, development organizations and the private sector.
www.cgiar.org

Front cover photo
Description.

Contributors
Tunrayo Alabi, Sika Gbegbelegbe, Vishnuvardhan Reddy Banda, and Dean Muungani

Characterization of Total Population Environments for cassava in Tanzania and banana in Ghana & Tanzania

Summary

This report presents Total Population Environments (TPEs) for cassava in Tanzania and plantain in both Tanzania and Ghana. The TPEs were developed using various environmental and socio-economic data. For cassava in Tanzania, five TPEs were developed. For bananas in Tanzania, 3 major TPEs were identified, whereas for banana in Ghana, 5 main TPEs were identified.

Table of Contents

Characterization of Total Population Environments for cassava in Tanzania and banana in Ghana & Tanzania

1 Characterization of Cassava production system in Tanzania to guide breeding efforts using spatial multivariate cluster and environmental similarity analysis

A better understanding of target environments is essential for Cassava breeding efforts for product development and varietal testing to optimally represent the target set of production environments (TPEs). Hence, in this activity, spatial tools were used to characterise the Cassava production environments into homologous mega-environments, having operational significance for breeding research. The analysis utilized spatial data on socioeconomic, climatic, soil and remote sensing vegetation indices and terrain attributes. The activity provides quantifiable information on where and how the Cassava crop is cultivated and under what conditions. The results of this analysis can guide the design of Product profiles to include specific traits and steer variety deployment to the appropriate agroecology.

1.1 Materials and methods

1.1.1 Acquisition and processing of Environmental and socioeconomic data

Developing a reliable TPE system for crop breeding heavily relies on the quality of climatic, soil, and terrain data. To achieve this goal, relevant and recent biophysical variables were gathered from various sources and thoroughly analyzed. Table 1 outlines these variables, which will be briefly discussed in this section.

To extract the number of rainy days, daily precipitation data from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) at a spatial resolution of 0.05° (approximately 5 km) for 1981-2021 were utilized. The amount of rainfall, its distribution, and intensities are crucial factors in crop development. Abiotic factors such as drought, heat stress, and the length of the dry season can negatively impact Cassava production, increasing the risk of crop failure in African countries. Therefore, annual drought frequencies were analyzed using long-term WORLDCLIM monthly data from 1961-2020.

Historical monthly maximum and minimum temperatures from 1961-2018, with a spatial resolution of 1km, were collected from WorldClim version 2.1. Additionally, the long-term value for mean sunshine hours from 1983-2015 was obtained from the EUMETSAT's Satellite Application Facility on Climate Monitoring (CM SAF) monthly data. Soil properties at a depth of 0-20cm, including soil organic carbon (SOC), total Nitrogen, available phosphorus, extractable

potassium, and soil pH, were analyzed using data from the recently produced ISDA soil properties map of Africa at a spatial resolution of 30m.

Remote sensing variables such as blue, red, green, and mid-infra-red reflectance bands and Normalized Difference vegetation indices (NDVI), were also used in the TPE analysis. These variables model the complex interaction of soil, climatic, terrain, and hydrologic features. Data from the Moderate Resolution Imaging Spectroradiometer (MODIS), obtained from NASA's Terra and Aqua satellites, were used to analyze land surface data of long-term mean monthly composites of about 18 years (2000 to 2017).

Terrain data from the Shuttle Radar Topography Mission (SRTM), which influences soil moisture distribution, soil erosion, and crop nutrient availability, were acquired from the USGS at a spatial resolution of 250 m. Socioeconomic variables, such as gridded human population data, subnational poverty levels, and crop production statistics were also analyzed. Current and future human population estimates were obtained from high-resolution African population projections from radiative forcing and socioeconomic models, 2000 to 2100. Poverty incidence data for 2016 were obtained from the International Poverty Line at Subnational Poverty levels from the World Bank, 2019, while crop production statistics data were from the Spatial Production Allocation Model (SPAM) 2017 v2.1 for Africa.

All data were pre-processed, and the coordinate system was standardized to the geographic reference system with a spatial resolution of 1 km. The details of the climatic, soil, and socioeconomic variables used in the TPE analysis are presented in Table 1.

Table 1: Environmental data used in the clustering for the development of TPEs

Data	Year	Source	Reference
Monthly rainfall (WORLDCLIM)	1961-2018	WorldClim 2.1	Fick & Hijmans 2017
Monthly rainfall (CHIRPS)	1981-2021	CHIRPS	Funk ***et al***. 2015
Daily rainfall data (CHIRPS)	1981-2021	CHIRPS	Funk ***et al***. 2015
Annual Rainy days (CHIRPS)	1981-2021	CHIRPS	Funk ***et al***. 2015
Drought frequency	1961-2020	WorldClim 2.1	Fick & Hijmans 2017
Monthly Max temperature	1961-2018	WorldClim 2.1	Fick & Hijmans 2017
Monthly Min temperature	1961-2018	WorldClim 2.1	Fick & Hijmans 2017
Normalized Vegetation Index (NDVI)	2000-2017	MODIS	AfSIS, 2017
Blue Reflectance	2000-2017	MODIS	AfSIS, 2017
Red Reflectance	2000-2017	MODIS	AfSIS, 2017
Green Reflectance	2000-2017	MODIS	AfSIS, 2017
Mid-Infrared Reflectance	2000-2017	MODIS	AfSIS, 2017
Mean monthly sunshine hour	1983-2015	CM SAF	Kothe et al., 2017
Human population data	2000-2100	DataGuru	Boke-Olén et al, 2017
Global subnational poverty data	2009-2016	World Bank	World Bank, 2019
Crop production data	2023	SPAM 2020	IFPRI,2023

Elevation Shuttle Radar Topography Mission	2013	USGS	SRTM 2013
Organic Carbon	2021	iSDAsoil	Hengl *et al.2021*
Soil pH	2021	iSDAsoil	Hengl *et al.2021*
Nitrogen	2021	iSDAsoil	Hengl *et al.2021*
Available Phosphorus	2021	iSDAsoil	Hengl *et al.2021*
Available Potassium	2021	iSDAsoil	Hengl *et al.2021*

1.1.2 Delineation of Cassava growing area in Tanzania

To determine the spatial extent of cassava TPE analysis, we employed the cassava suitability data from GAEZ version 4, which was created by the Food and Agriculture Organization of the United Nations (FAO) and the International Institute for Applied Systems Analysis (IIASA) (as shown in Figure 2). Nearly all the areas of Tanzania have suitability levels for cassava ranging from marginal to high suitability (Figures 1 & 2).

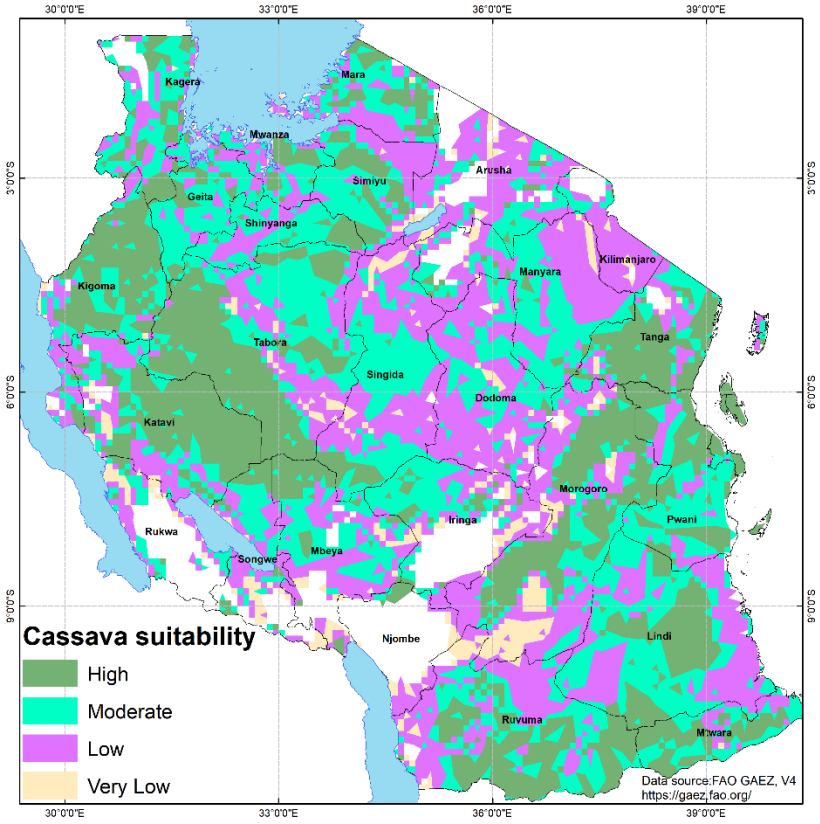


Figure 1: Cassava suitability using soil and climatic characteristics

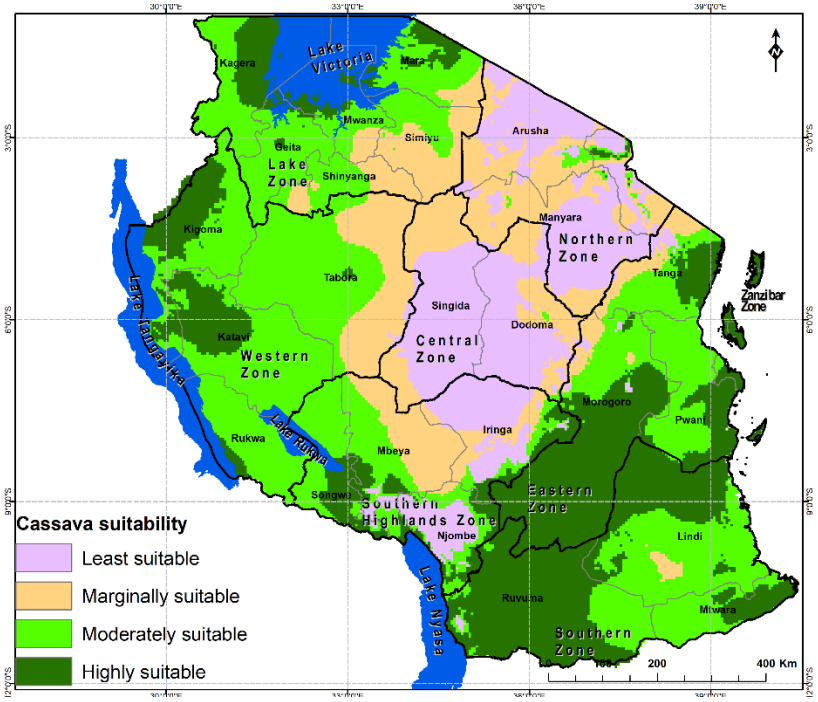


Figure 2: Cassava suitability using climatic characteristics

1.1.3 Principal component analysis

In data analysis, Principal Component Analysis (PCA) is a popular method for reducing dimensionality. It involves transforming data into an orthogonal space to eliminate redundancy and autocorrelation, while still capturing the original dataset's variability in fewer layers. By extracting the maximal variance in the data space, PCA helps to capture the most variability from historical rainfall and minimum and maximum temperature data.

To avoid losing important trends and variations in the data, monthly data was used instead of computing the mean over the years. Typically, only the principal components (PCs) that captured over 90% of the total variance were utilized in the Cassava TPE development analysis. For instance, in the TPE clustering analysis, fourteen PCs which accounted for about 98% of the total variance of the monthly rainfall were employed, while nine PCs with 98.8% variance were applied for the drought frequency data. Moreover, five PC layers accounting for 98.4% of the monthly maximum and minimum temperatures were utilized in the TPE development. By identifying and utilizing only the most significant principal components, we effectively captured and analyzed the most important factors in our data. Overall, the PCA method is an effective way to reduce data dimensions while still retaining vital information.

1.1.3.1 Development of Cassava Total Population Environments (TPEs) in Tanzania

A better understanding of the environmental context in the target region is essential to unlocking the Cassava potential for food security and wealth creation. We utilized a rich resource and diverse

arrays of environmental data in the form of climatic, edaphic, topographic, and remote sensing layers (Table 1) in a spatial multivariate clustering analysis for the target area. We used the Kmeans clustering technique, an unsupervised learning algorithm, to solve the clustering problems in machine learning or data science. It is an iterative algorithm that divides the unlabelled dataset into k different clusters so that each dataset belongs to only one group with similar properties. Kmeans algorithm was performed with hyperparameter tuning to find the optimal number of clusters. Kmeans algorithm was trained with the clValid package in R. The R package clValid contains functions for validating the results of clustering analysis and helps to help determine the most appropriate method and the number of clusters for the environmental datasets. This resulted in statistically independent homologous clusters or target populations of the environment (TPE) to guide and operationalize breeding research in the target countries. The mega-environment map generated can aid in scaling up and out of project outcomes for wider impact. The steps and procedures for the development of TPEs are shown in Figure 3.

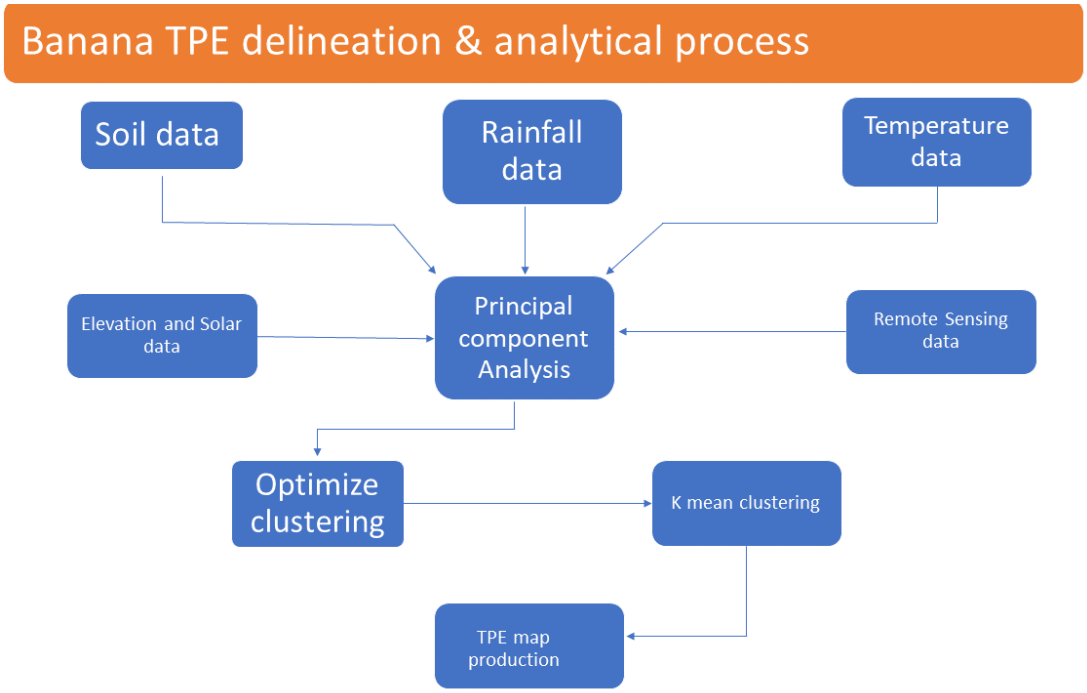


Figure 3: The Analytical workflow for Cassava TPEs development

1.1.3.2 Socioeconomic analysis of Cassava's TPEs to estimate the potential impacts

Spatial analysis was also performed to examine socioeconomic contexts and the potential impact of Cassava product development. High agricultural potential areas characterised by dense rural population, significant poverty levels, and high market accessibility could be of higher priority for demonstrating the impact of improved agricultural technologies. The potential impact of improved Cassava technologies was calculated using five socioeconomic variables: total human population,

poverty levels, the projected total population in 2050, Cassava production, and malnutrition data of stunting.

1.1.4 Results

Tanzania's clustering analysis yielded five main TPEs (Figure 4). The main cassava growing zones comprise southern, eastern and lake zones. Some Cassava-growing areas are also found in part of the western and south highland zones.

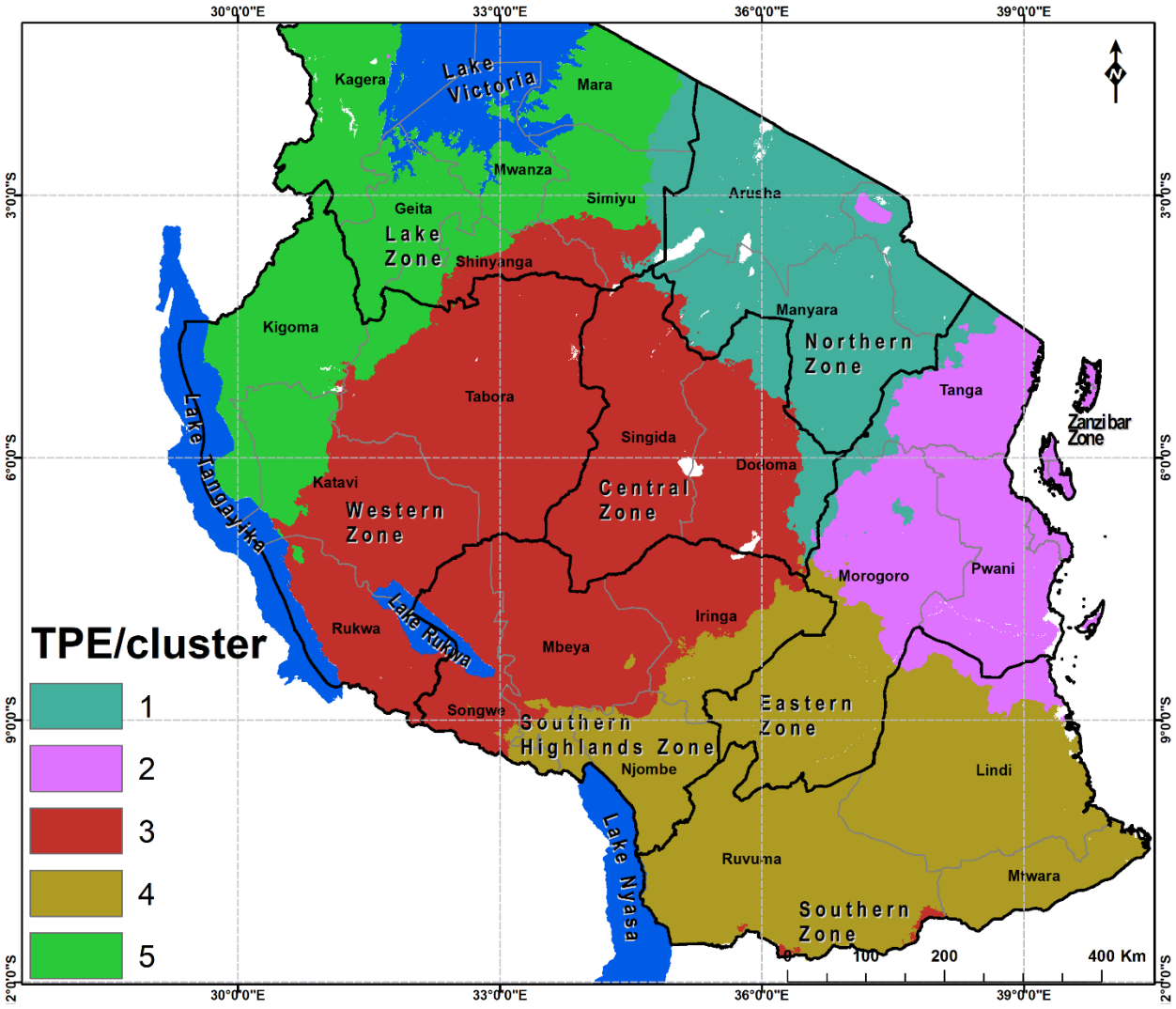


Figure 4: Map of Tanzania showing 5 clusters/TPEs for cassava

Table 2: Cassava production characteristics of the Target Population of the Environment (TPEs)

TPE	Cassava area harvested (ha)	Cassava production (tons)	Cassava Yield (ton/ha)	% Cassava area harvested	% Cassava production
TPE1	43,350	436,155	10.1	4.50	6.30
TPE2	137,078	1,308,616	9.5	14.22	18.91
TPE3	110,124	994,349	9.0	11.42	14.37
TPE4	236,095	1,405,418	6.0	24.49	20.31
TPE5	437,389	2,774,279	6.3	45.37	40.10

Table 3: Population and poverty characteristics of the TPEs

TPE	Total Population 2020	Total Population 2050	Urban population 2020	Rural population 2050	Number of urban pop in poverty	Number of rural pop in poverty	Number of urban pop under nourished	Number of rural pop undernourished
TPE1	8,834,775	26,827,425	4,818,302	11,121,802	3,895,244	2,303,201	3,741,055	4,785,228
TPE2	13,933,953	52,552,692	3,951,790	9,766,585	4,983,341	1,626,094	6,781,122	5,172,237
TPE3	14,387,977	37,132,675	10,135,440	20,613,464	8,057,984	5,747,548	5,971,351	8,393,603
TPE4	8,355,284	19,940,624	6,301,283	12,881,764	3,776,174	3,005,806	4,647,875	7,408,938
TPE5	17,344,855	48,768,186	10,561,754	23,708,692	11,137,211	6,773,444	6,548,884	9,372,276

2 Characterization of Banana production system in Ghana and Tanzania to guide breeding efforts using spatial multivariate cluster and environmental similarity analysis

A better understanding of target environments is essential for banana breeding efforts for product development and varietal testing to optimally represent the target set of production environments (TPEs). Hence in this activity, spatial tools were used to characterize the banana production environments into homologous mega-environments, having operational significance for breeding research. The analysis utilized spatial data on socioeconomic, climatic, soil and remote sensing vegetation indices and terrain attributes. The activity provides quantifiable information on where and how the Banana crop is cultivated and under what conditions. The results of this analysis can guide the design of Product profiles to include specific traits and steer variety deployment to the appropriate agroecology.

2.1 Materials and methods

2.1.1 Acquisition and processing of Environmental and socioeconomic data

Varieties of climatic data from the recently released daily and monthly CHIRPS and WorldClim data (1960-2021) were downloaded, processed, and analyzed for the banana production environments in Africa. ISRIC soil fertility, depth, texture, and acidity data at a spatial resolution of 250m were also downloaded and processed for analysis. Socioeconomic data such as human population density, market access, and poverty incidence were acquired and processed for the environmental analysis. Gridded Banana production data was also obtained and employed in the ecological characterization analysis.

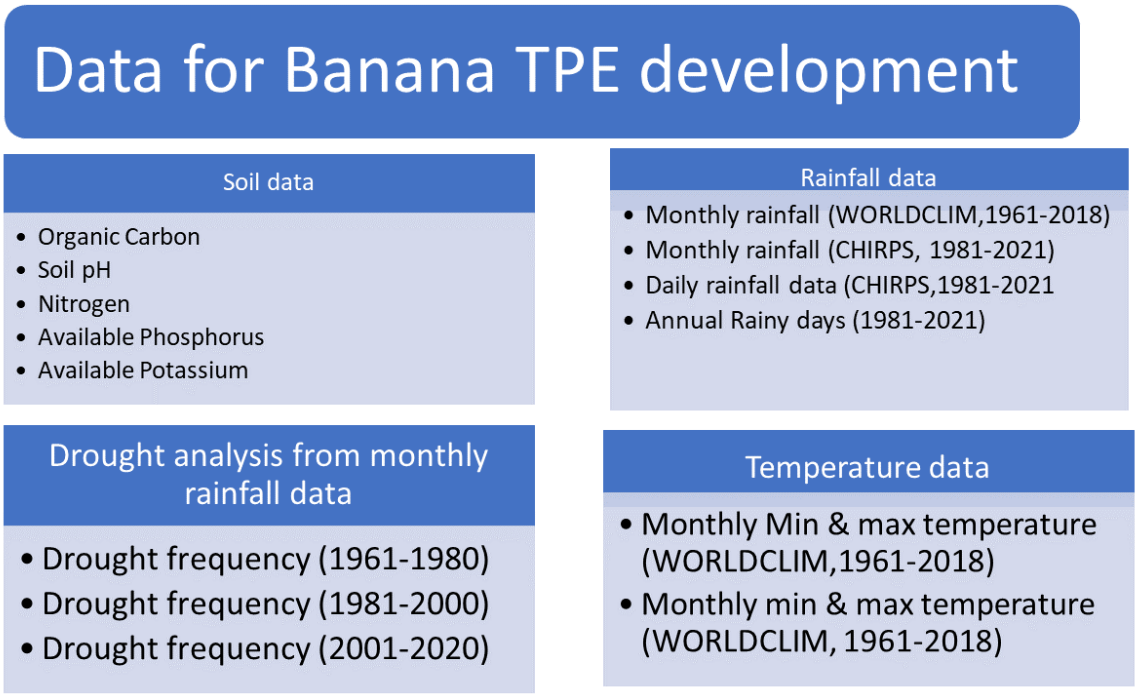


Figure 1a: Environmental data used in the clustering for the development of TPEs

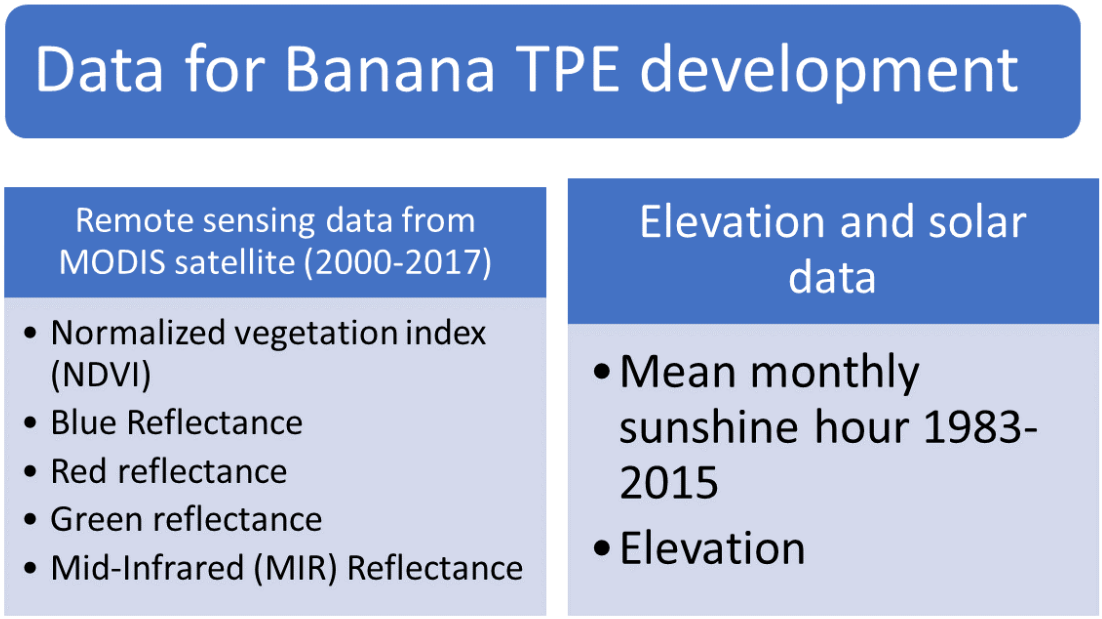


Figure 1b: Environmental data used in the clustering for the development of TPEs

2.1.2 Development of Banana mega-environments in the two target countries, Ghana and Tanzania

A better understanding of the environmental context in the target region is essential to unlocking the banana/plantain potential for food security and wealth creation. We utilized a rich resource and diverse arrays of environmental data in the form of climatic, edaphic, topographic, and remote sensing layers (Figure 1a and 1b) in a spatial multivariate clustering analysis for the target area. We used the Kmeans clustering technique, an unsupervised learning algorithm, to solve the clustering problems in machine learning or data science. It is an iterative algorithm that divides the unlabeled dataset into k different clusters so that each dataset belongs to only one group with similar properties. Kmeans algorithm was performed with hyperparameter tuning to find the optimal number of clusters. Kmeans algorithm was trained with the clValid package in R. The R package clValid contains functions for validating the results of clustering analysis and helps to help determine the most appropriate method and the number of clusters for the environmental datasets. This resulted in statistically independent homologous clusters or target populations of the environment (TPE) to guide and operationalize breeding research in the target countries. The mega-environment map generated can aid in scaling up and out of project outcomes for wider impact. The steps and procedures for the development of TPEs are shown in Figure 3.

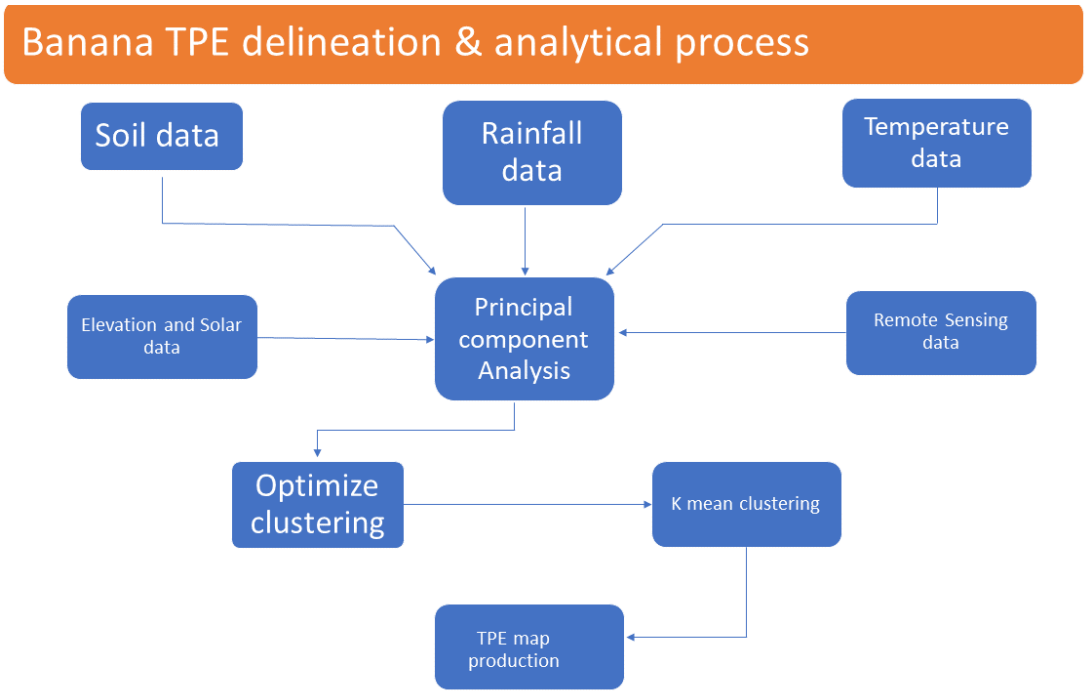


Figure 3: The Analytical workflow for Banana TPEs development

2.1.3 Socioeconomic analysis of the banana growing environment to estimate the impact

Spatial analysis was also performed to examine socioeconomic contexts and the potential impact of banana product development. High agricultural potential areas characterized by the dense rural population, significant poverty levels, and high market accessibility could be of higher priority for demonstrating the impact of improved agricultural technologies. The potential impact of improved banana technologies was calculated using five socioeconomic variables: total human population, poverty levels, the projected total population in 2050, banana production, and market accessibility.

2.2 Results

The clustering analysis results are presented in Figures 4-6 for Ghana. Figure 4 shows the optimal numbers of clusters for Ghana using environmental layers. Ghana was divided into five main TPEs. The southern forest zone was divided into the coastal TPE, western TPE and Eastern TPE around Volta Lake. TPE 3 covered mainly the Northern region of Ghana, while TPE 5 covered part of the northern region, upper west and upper east regions. Further subdivisions of these TPEs are shown in Figures 5 and 6. As more TPEs were delineated, the banana-growing area of Ghana has more diversities than the northern part.

Similarly, Tanzania's clustering analysis yielded five main TPEs (Figure 7). Banana growing zones comprise southern, eastern and lake zones. Some banana-growing areas are also found in part of the western and south highland zones. Further subdivisions of the TPEs are shown in figures 8 and 9.

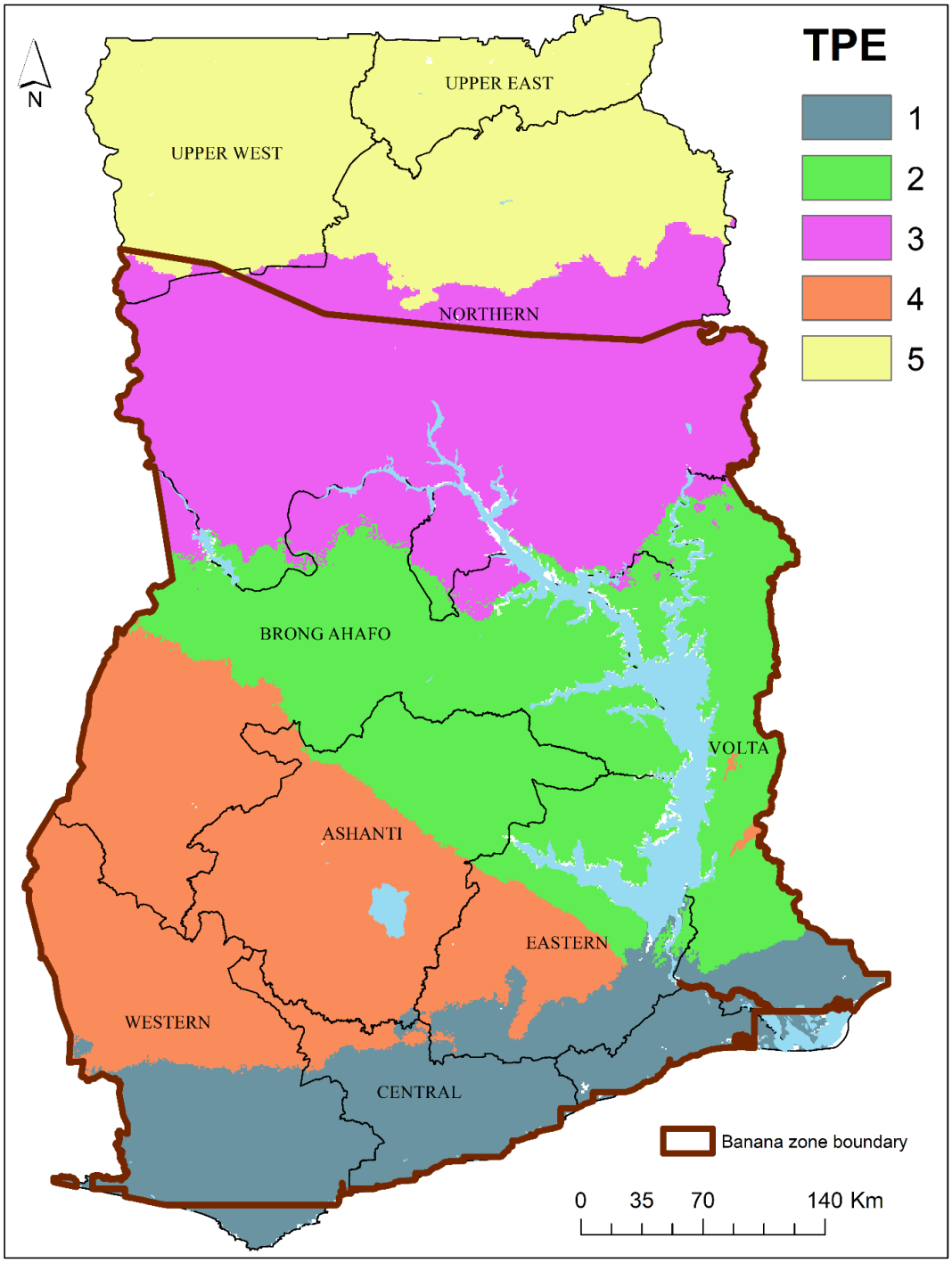


Figure 4: Optimal number of clusters showing five main TPEs in Ghana. Note that 4 of the TPEs are within the banana-growing area.

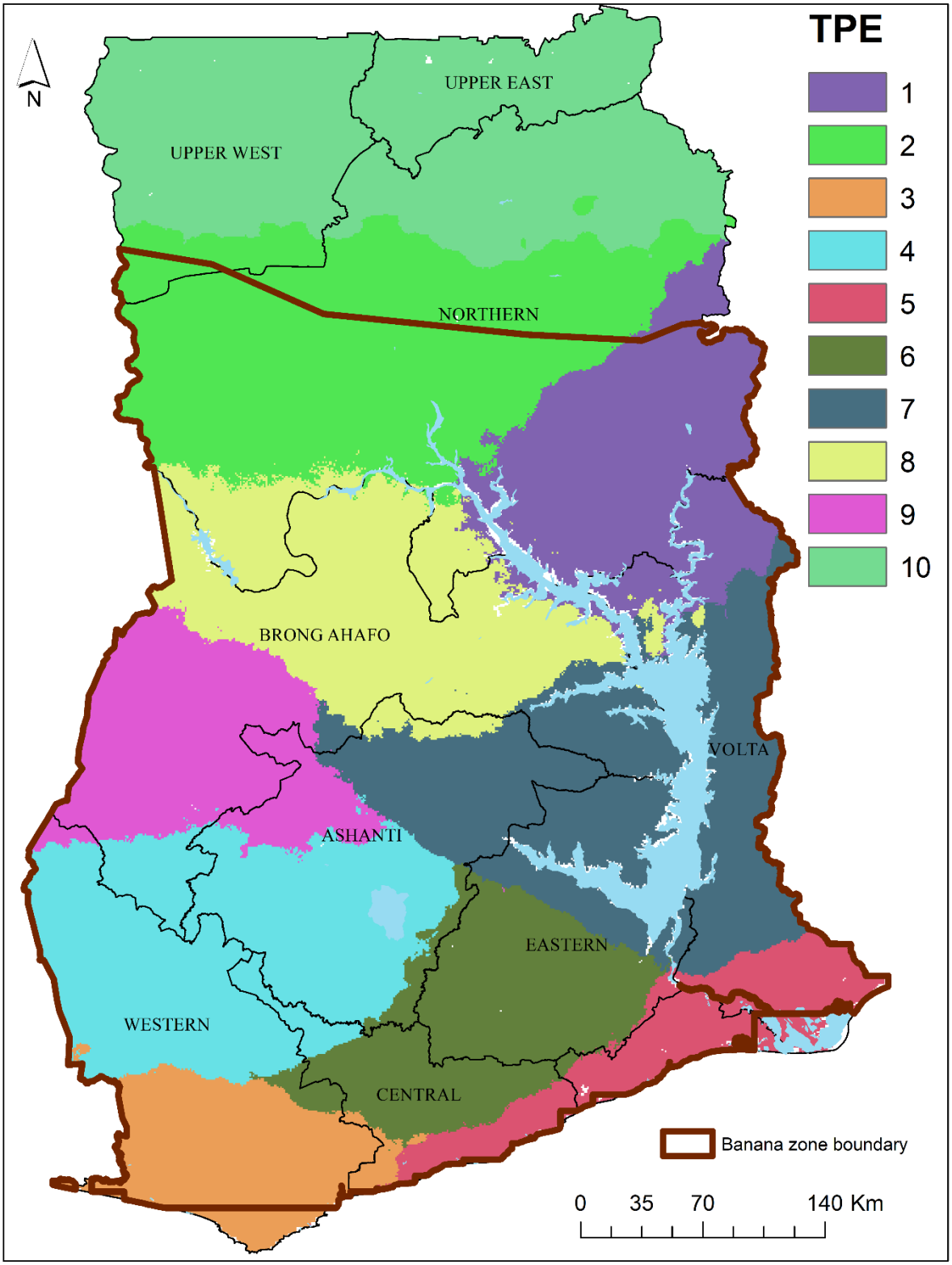


Figure 5: Map of Ghana showing 10 clusters/TPEs. Note that 9 TPEs are within the banana growing zone

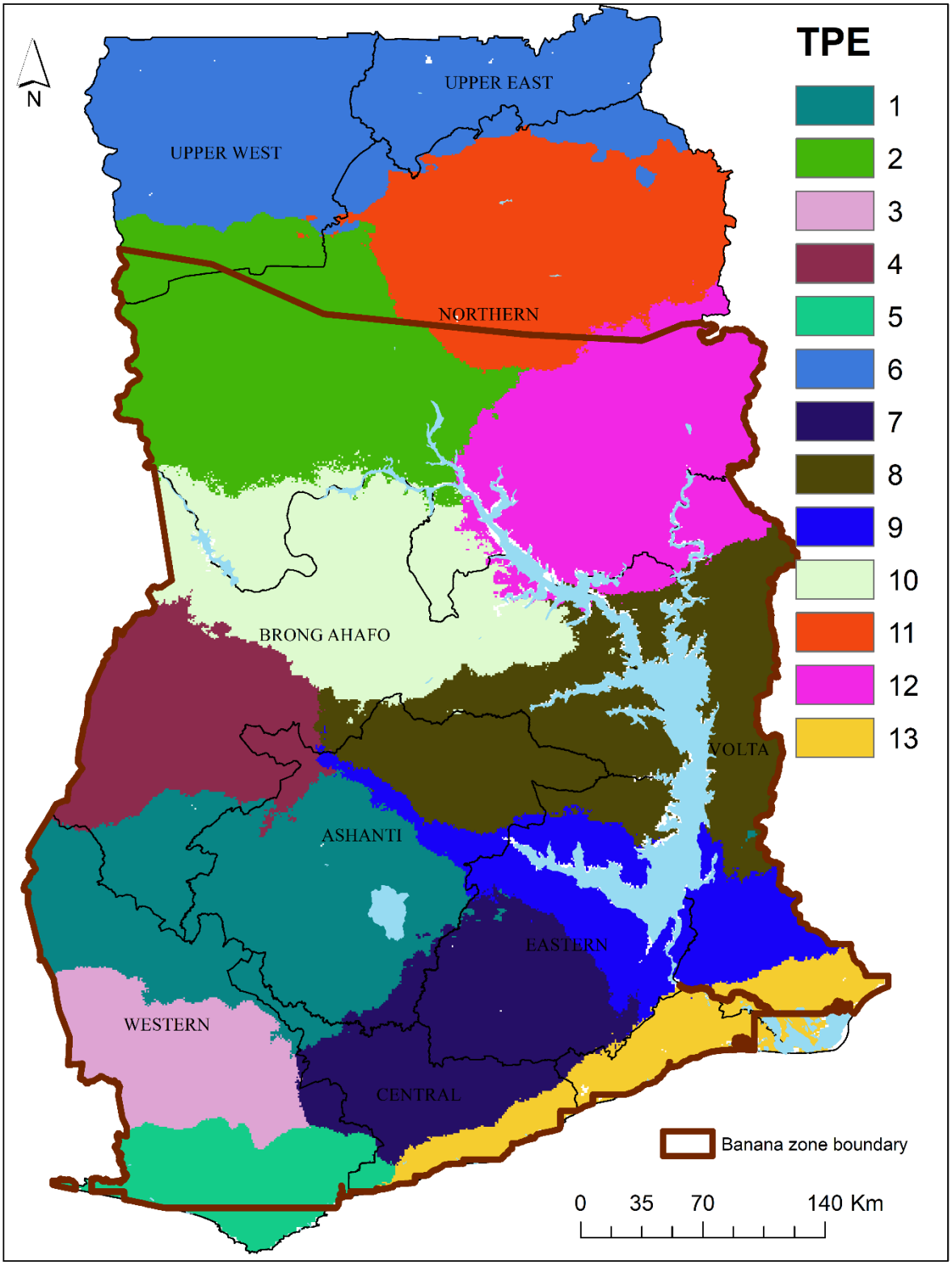


Figure 6: Map of Ghana showing 13 clusters/TPEs. Note that 11 TPEs are within the banana growing zone

Figure 7: Map of Tanzania showing 5 clusters/TPEs. Note that 3 TPEs are entirely within the banana growing zone

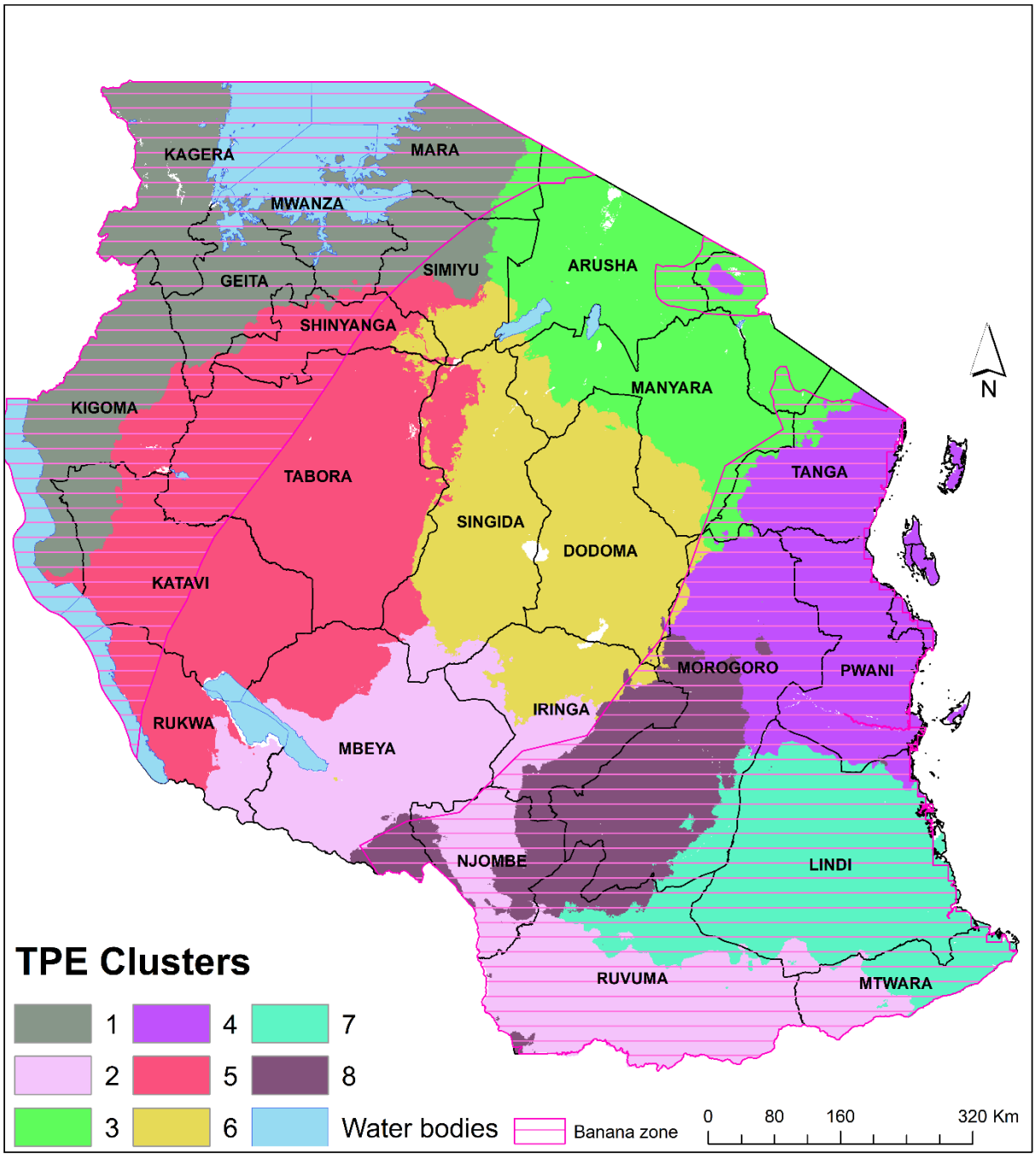


Figure 8: Map of Tanzania showing 8 clusters/TPEs. Note that 6 TPEs are entirely within the banana growing zone

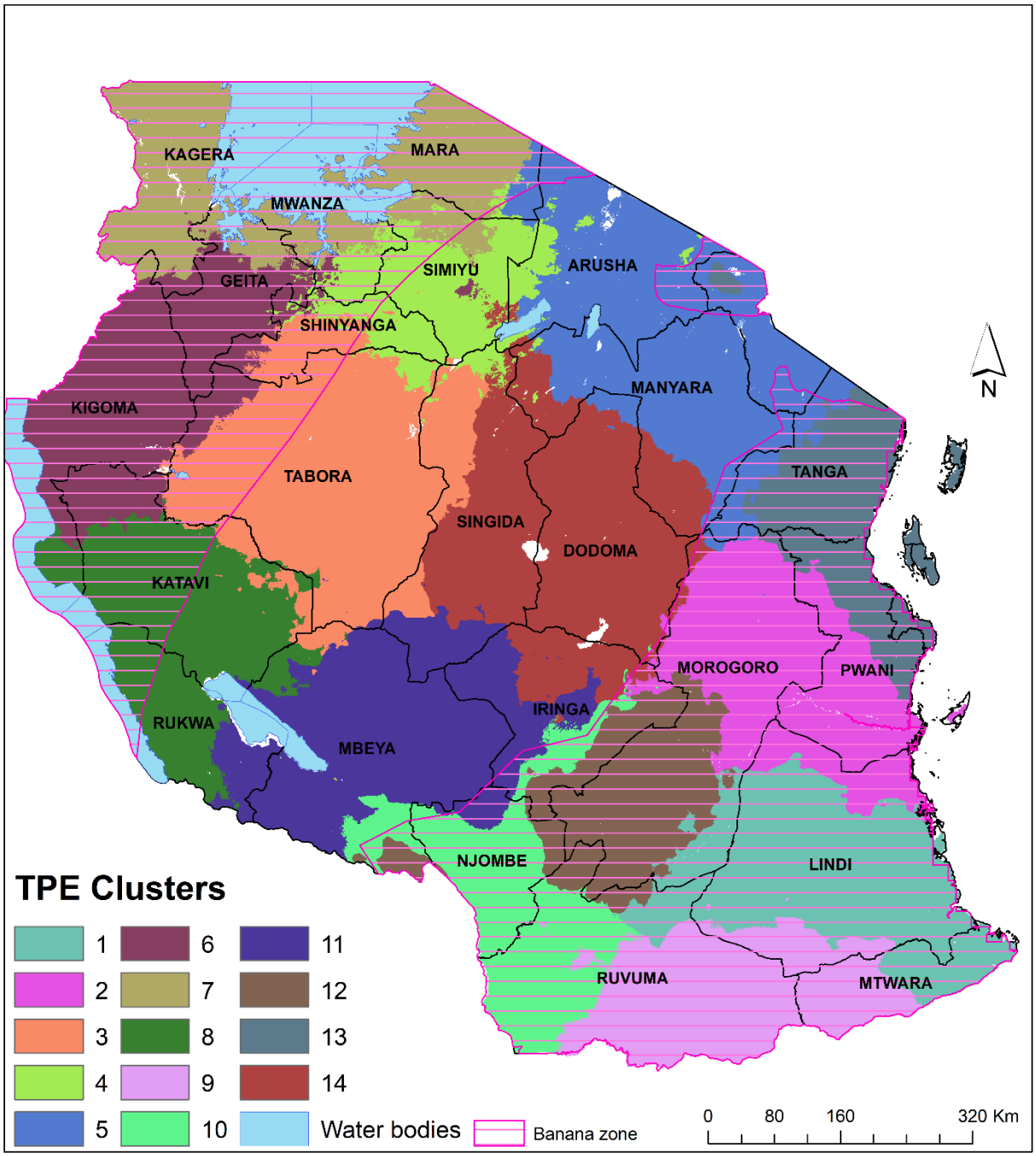


Figure 9: Map of Tanzania showing 14 clusters/TPEs. Note that about 9 TPEs are entirely within the banana growing zone

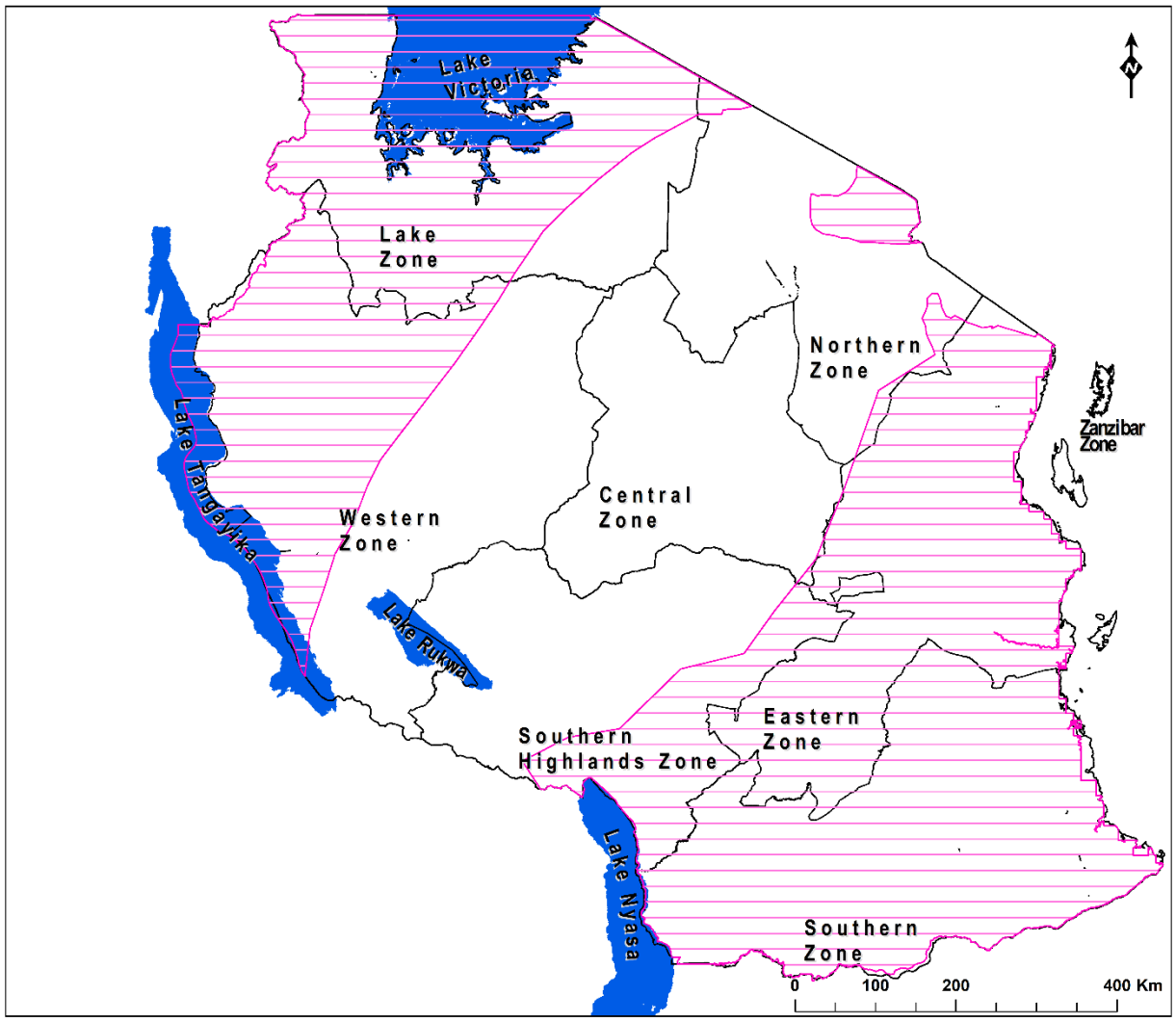


Figure 10: Map of Tanzania showing Agricultural zones