ANNEX B Geospatial data processing

To create a typology of the rural opportunity space (ROS), the *Rural Development Report 2019* uses commercialization potential and agricultural production potential as the two axes of the opportunity space (Wiggins and Proctor, 2001; Ripoll et al., 2017).[52] Commercialization potential is proxied by population density data drawn from the WorldPop project, while agricultural potential is proxied by the Enhanced Vegetation Index (EVI), which is based on satellite observations.[53] The ROS is employed at both the global level and the household level in the report. The details of how geospatial data were processed for these analyses are presented below.

Global ROS analyses

The WorldPop project was launched in July 2009 with the aim of producing detailed and freely available population distribution maps. It provides 1-km and 100-m spatial resolution (i.e. the edge length of a single grid cell) population density maps (number of people per grid cell) for each country in sub-Saharan Africa (SSA), Latin America and the Caribbean (LAC), and Asia and the Pacific (APR). The production of the WorldPop spatial datasets generally follows the methodologies outlined in Tatem et al., 2007; Gaughan et al., 2013; Alegana et al., 2015; and Stevens et al., 2015.[54, 55] In most countries, population estimations exist for two epochs, namely 2010 and 2015. WorldPop also includes age- and gender-differentiated spatially explicit information on population distributions, albeit these are at a 1-km spatial resolution (unlike the population densities, which are also provided at a 100-m resolution). Given the 2019 report's focus on age- and gender-differentiated distributions of rural youth over the ROS, data with a 1-km resolution were used to define the commercialization potential using population densities over the rural-urban gradient at the global level.

In order to define a globally comparable scale, all numeric values of the grids in the data were ordered from least to most dense, and population was successively summed to create four groups (quartiles) with populations of equal size, ranging from the least to the most densely settled areas, to create a rural-urban gradient. The least dense quartile corresponds to rural areas and the densest quartile to urban areas. In between are the semi-rural (second quartile) and peri-urban (third quartile) areas. The bottom three population density quartiles (rural, semi-rural and peri-urban areas) are referred to as rural (i.e. non-urban) in this report. Rural, semi-rural and peri-urban areas, respectively, represent the categories of *low, medium and high commercial potential* on the vertical axis of the global ROS.

For the agricultural potential axis of the ROS, Moderate Resolution Imaging Spectroradiometer Enhanced Vegetation Index (MODIS-EVI) grids with a 250-m resolution were resampled to 1 km using a nearest neighbour algorithm to match the resolution of age- and gender-disaggregated WorldPop grids. Several pre-processing steps reduced the effects of residual clouds and shadows, dust, aerosols, off-nadir viewing and low sun zenith angles in the EVI data. First, pixels that were flagged as no data, snow/ice or cloud in the MOD13Q1 pixel reliability layer prior to filtering, based on MODIS quality assurance information, were excluded. Only pixels labelled as "good data" or "marginal data" were retained, i.e. pixels in the quality layer that were flagged as either zero or one. Second, data gaps were linearly interpolated. Third, the time series were smoothed using the Savitzky-Golay approach (Chen et al., 2016).[56] Fourth, EVI values were used only for land classified as cropland or pastureland. To do so, a new global map of cropland/ pastures was created by fusing two existing maps (Waldner

52 S. Wiggins and S. Proctor, 2001. How Special Are Rural Areas? The economic implications of location for rural development. *Development Policy Review*, 19 (4): 427-436; S. Ripoll, J. Andersson, L. Badstue, M. Büttner, J. Chamberlin, O. Erenstein and J. Sumberg, 2017. Rural Transformation, Cereals and Youth in Africa: What role for international agricultural research? *Outlook on Agriculture*, 46 (3): 1-10.

53 https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod13q1_v006_.

54 A.J. Tatem, A.M. Noor, C. von Hagen, A. Di Grigorio and S.I. Hay, 2007. High-Resolution Population Maps for Low-Income Nations: Combining land cover and census in East Africa. *PLoS One*, 2: 34-36; A.E. Gaughan, F.R. Stevens, C. Linard, P. Jia and A.J. Tatem, 2013. High-Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. *PLoS One*, 8.; V.A. Alegana, P.M. Atkinson, C. Pezzulo, 2015. Fine Resolution Mapping of Population Age-Structures for Health and Development Applications. *Journal of the Royal Society Interface*, 12:20150073-20150073.; F.R. Stevens, A.E. Gaughan, C. Linard and A.J. Tatem, 2015. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLoS One*, 10:e0107042.

55 For further details and publications, see: http://www.worldpop.org.uk/data/methods/.

56 J. Chen, P. Jönsson, M. Tamura, Z. Gu, B. Matsushita and L. Eklundh, 2004. A Simple Method for Reconstructing a High-Quality NDVI Time-Series Data Set Based on the Savitzky-Golay Filter. *Remote Sensing of Environment*, 91: 332-344 (doi:10.1016/j.rse.2004.03.014).

et al., 2016; GFSAD, 2010).[57] By doing so, the analysis was spatially targeted at agricultural land, and production potential could then be proxied. Finally, average EVI values for the three-year period between 2013 and 2015 were alculated to avoid seasonality and agroclimatic variation. EVI grids (the same as the WorldPop grids) were ordered from lowest to highest, with each of the three groups (terciles) containing one third of all the non-urban space and representing the categories of *low, medium and high agricultural potential* on the horizontal axis.

Using the above data, the number and share of rural (non-urban) youth in each of the ROS categories were calculated, and these data were then used in the analyses presented throughout the report.

Household-level ROS analyses

The household-level data used in the report cover over 765,000 individuals (128,227 of whom were classified as rural youth, representing around 134 million young people in rural areas) in 12 countries across 3 regions (SSA, APR and LAC). These data were used to analyse the ways in which these young people and their families are engaging with the economy and to position them in the ROS using the available georeferenced information about the administrative layer with the highest spatial resolution (this varied across surveys).

Household data for sub-Saharan Africa (SSA) were drawn from the Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) data, all of which provide georeferenced information (i.e. the centroids) for each enumeration area (EA) in the sample. Using the number of dwellings and the average household size in each EA, the total population of the average EA in each country was calculated. Using the known population distribution from the WorldPop data, a boundary was then drawn around the EA centroid to capture a population of this size, which created artificial EA boundaries. The population density of each EA was calculated and then classified on the basis of the population density quartiles along the rural-urban gradient (using the global threshold defined above): Q1=Rural, Q2=Semi-Rural, Q3=Peri-Urban, Q4=Urban. The artificial boundaries created for each EA were also used to calculate the average value of the EVI for the 2013-2015 period (because most LSMS-ISA data were collected in 2014 or 2015), as described above.

For APR and LAC, household data sources do not include georeferenced information. Centroids of municipalities or other small administrative units and, in some cases, boundaries (polygons) for relatively small administrative areas obtained from the DIVA-GIS database were therefore used to repeat the above process to create the ROS variables for each household.

Household-level information was combined with the above georeferenced information in order to assign each household (and the young people within them) to one of the categories of the ROS as defined in figure 2.4. This information was then used in analyses designed to afford a fuller understanding of how the ROS affects school-to-work transitions and the labour force participation rates of youth versus adults (along with gender differentiation).

57 F. Waldner, S. Fritz, A. Di Gregorio, D. Plotnikov, S. Bartalev, N. Kussul, P. Gong et al., 2016. A Unified Cropland Layer at 250 m for Global Agriculture Monitoring. *Data*, 1:3 (doi:10.3390/data1010003); Global Food Security-Support Analysis Data (GFSAD): https://www.usgs.gov/centers/wgsc/science/global-food-security-support-analysis-data-30-m?qt-science_center_objects=0#qt-science_center_objects.