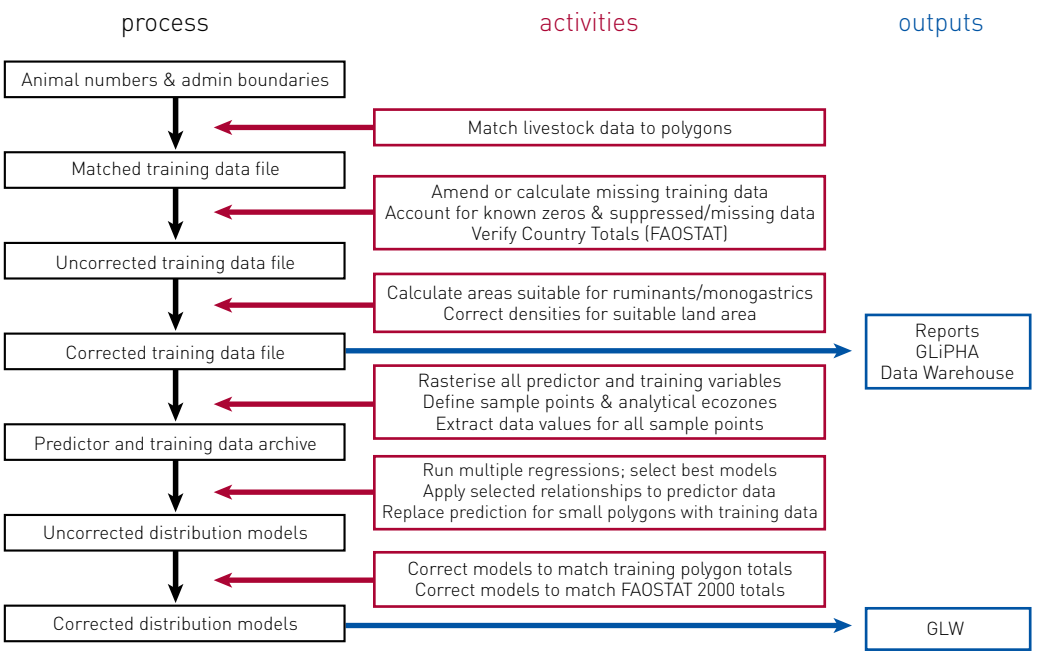
# 5 Modelling livestock distribution

Once the available agricultural statistics have been collected, standardized, enhanced with supplementary data and adjusted for the extent of land deemed suitable for livestock production, the resulting data archive provides a sound basis for statistical distribution modelling. This process depends on establishing a robust statistical relationship between livestock numbers and one, or more, predictor variable for which data are available for the entire area of interest. These relationships are detailed later in this section.

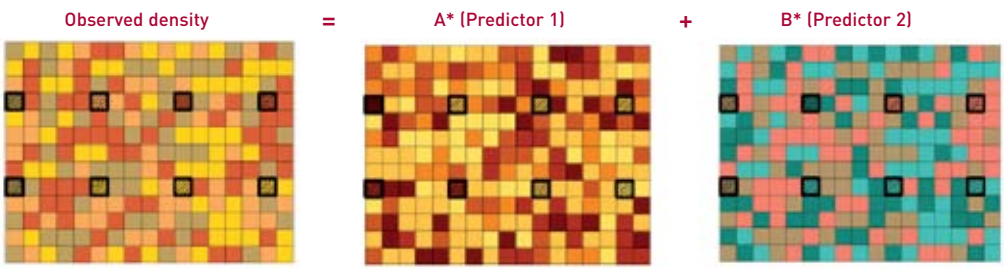
The modelling process, including inputs and outputs at the various stages, is summarized in Figure 5.1. This process relies on the use of raster images to store both observed (or training) data (i.e. livestock densities) and all the predictor variables. Statistical relationships are established between observed livestock densities and predictor variables using values extracted for a series of regularly spaced sample points, as illustrated in Figure 5.2. The resulting equations are then applied to all pixels in the predictor images so as to produce a predicted distribution map.

As a result, the technique can be used to predict livestock densities in areas for which no livestock data are available, i.e. filling in gaps. Moreover, because predicted densities are produced at the resolution of the raster imagery, the models generate heterogeneous densities within polygons that have only one single observed value, thus disaggregating the original data. For limited datasets, therefore, the method has the major advantage of

5.1 SCHEMA OF MODELLING PROCESS



**5.2** SCHEMA OF SAMPLING PROCESS




1) Convert all data maps to images with the same pixel size (resolution);
2) Extract values for observed values of livestock density, and for each predictor variable at fixed sample points (hatched squares);
3) Calculate a regression equation of the form:
   *Observed density = Constant + A * (Predictor 1) + B * (Predictor 2) + ...*;
4) Providing the equation is statistically significant (i.e. reliable), apply the right hand side of the equation to **all** pixels in the predictor variable images to produce the predicted density;
5) Repeat the process for each of a series of analysis zones (e.g. ecozones).

both filling in gaps and refining the level of detail that can be mapped.

As the predictors of animal density are unlikely to be consistent from region to region, the modelling process should be run at several different spatial scales to provide a range of predictive relationships appropriate to specific areas. As well as administrative-level analyses an ecological stratification has been routinely used, on the assumption that the factors determining animal distributions are likely to be similar in areas with comparable ecological characteristics, thereby allowing (i) more robust statistical relationships between training data and predictor variables to be established, and (ii) more realistic predictions of livestock densities in other parts of the same ecological zone for which data are not available.

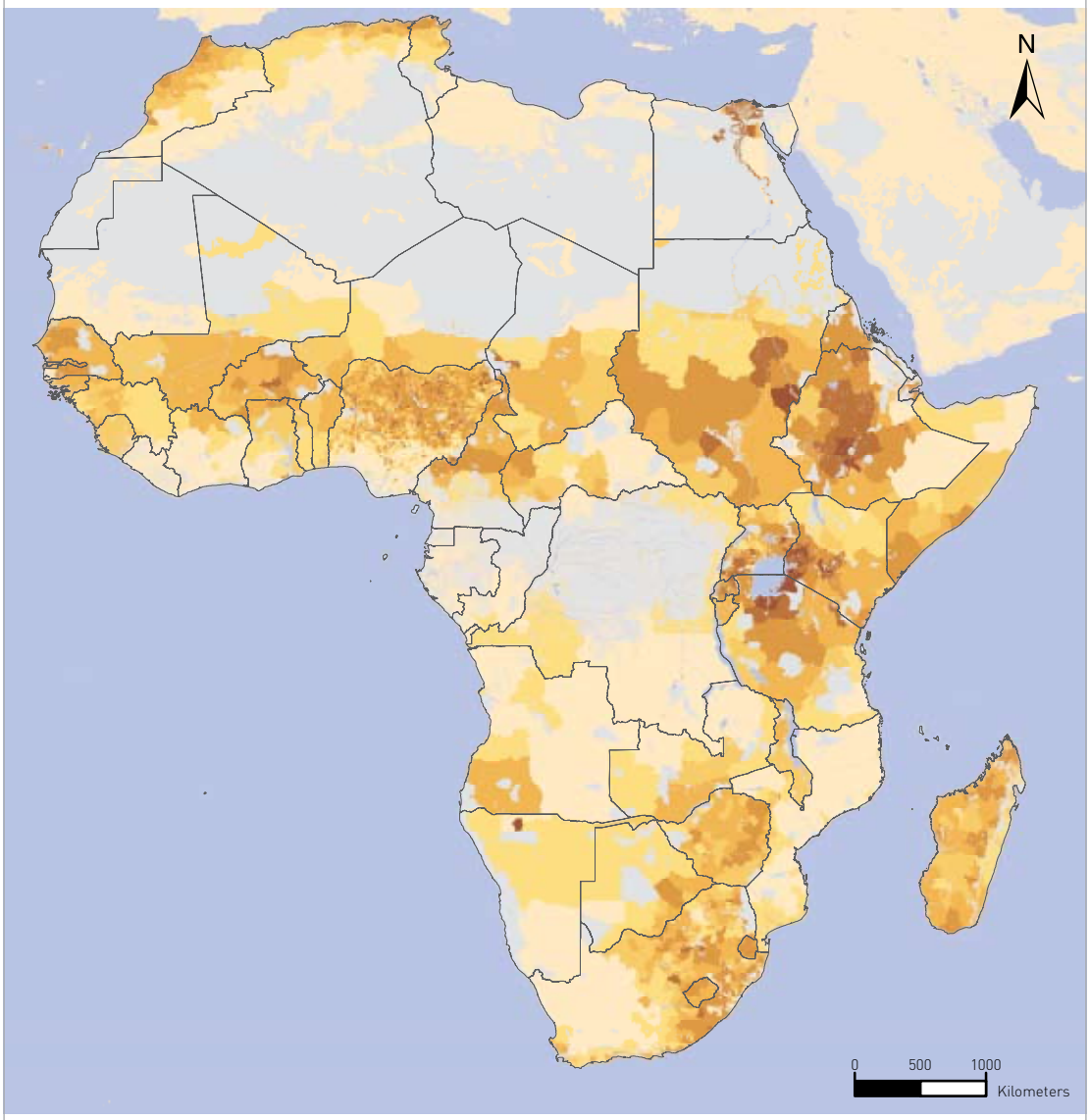
The ecological zones used to stratify the modelling were defined separately for each continent using non-hierarchical clustering techniques, either within the ADDAPIX programme (Griguolo and Mazzanti, 1996) or ERDAS Imagine software (Leica Geosystems®). The input parameters were drawn from the suite of predictor variables and included elevation and a series of remotely sensed parameters (the mean [Fourier component 0] and phase [Fourier component 1] of middle infrared, land surface temperature, vegetation index, air temperature and vapour pressure deficit). See below for further details.

## A WORKED EXAMPLE - AFRICA

The whole modelling process can be illustrated using cattle in Africa as an example. Observed cattle densities were derived from various national census reports, livestock surveys and data archives for the period 1992-2003. As can be seen in Figure 5.3, most known, or 'observed', cattle densities relate to administrative units, some of which are very large.

A series of stepwise multiple regression analyses was performed to establish the statistical relationships between observed cattle densities and a range of predictor variables drawn from those described below, including: satellite-derived measures of rainfall, temperature, vapour pressure deficit, vegetation cover and elevation (provided by the Trypanosomiasis And Land-use in Africa

**5.3** OBSERVED CATTLE DENSITIES IN AFRICA

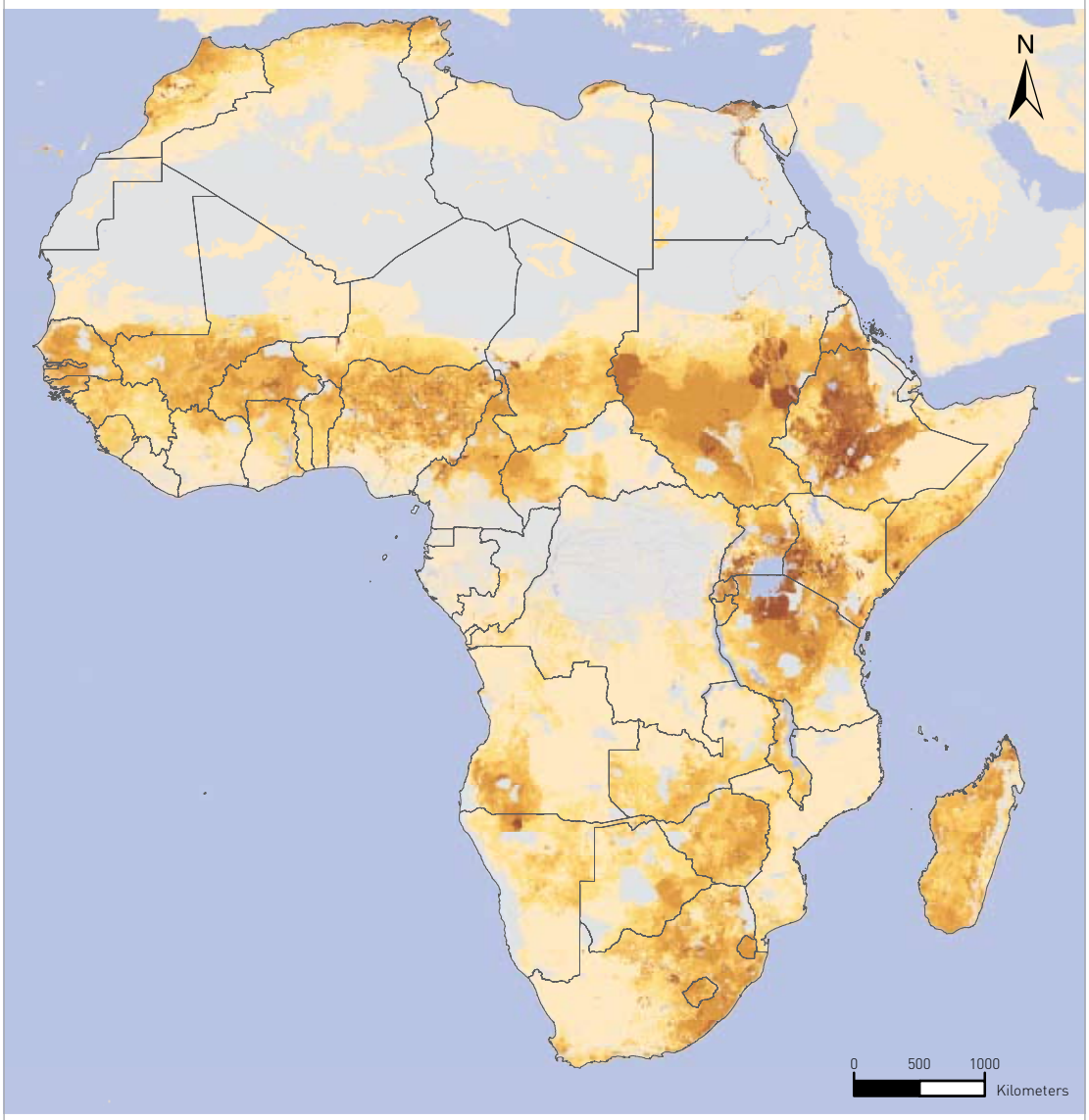



Number per square km

| | | | |
|---|---|---|---|
| <1 | 10–20 | 100–250 | International boundary |
| 1–5 | 20–50 | >250 | Water |
| 5–10 | 50–100 | | Unsuitable for ruminant |

*Source:* Robinson *et al.*, 2007.

**5.4** MODELLED CATTLE DENSITIES IN AFRICA




Number per square km

| | | | |
|---|---|---|---|
| <1 | 10–20 | 100–250 | International boundary |
| 1–5 | 20–50 | >250 | Water |
| 5–10 | 50–100 | | Unsuitable for ruminant |

*Source:* Robinson *et al.*, 2007.

(TALA) Research Group at the University of Oxford); potential evapotranspiration; length of growing period (LGP); human population; and the potential number of tsetse species present.

Values were extracted for approximately 19 000 regularly spaced sample points and a series of regression models derived at different spatial scales: (i) the entire continent; (ii) four continental subregions (east, west, south and north); (iii) 50 ecological zones; and (iv) each ecological zone within each country. In addition, four sets of transformations were assessed – no transformation, logarithmic, exponential and power – to address the possibility that relationships were non-linear. The best relationship was then selected based on $R^2$ values. Approximately 500 equations were generated and assessed, of which some 60 were selected to contribute to the model. Where statistically valid equations were not found at the highest spatial resolution, equations for the next spatial scale up (region) were assessed, and so on until an acceptable model was identified for each unit of area.

The resulting equations were then applied to the original imagery to generate a map of predicted cattle distribution at a spatial resolution of three minutes of arc (approximately 5 $km^2$ at the equator). All the predictive equations used were statistically significant at the 1 percent level ($p<0.01$), or better; but it is axiomatic that the validity of the predicted distribution map depends primarily on the accuracy of the 'observed' training data.

The predicted cattle distribution in Africa, as shown in Figure 5.4, mirrors the observed distribution (Figure 5.3) very well and picks out both major foci (e.g. the Kenya, Ethiopia and Zimbabwe highlands, Tanzania, semi-arid and dry subhumid West Africa), as well as smaller concentrations such as the Gezira irrigation scheme in Sudan, the inland delta of the Niger River in Mali and southeastern Zambia. Relatively high-resolution observed data for Nigeria, derived from aerial survey, were smoothed by the regression analysis. Some of the contrasts between observed and predicted maps are due to minor differences in values falling into different mapping classes. There are also some minor anomalies in northern Chad, where very high predicted densities are obviously false and are caused by extreme predictor values. Human population density is a major determinant of cattle distribution in Africa (Bourn and Wint, 1994) and was the primary predictor in 30 percent of regression equations used.

There is, of course, a danger that these predictive techniques, based on intensive processing, may conceal substantial errors; it is all too easy to be seduced by the fact that a somewhat messy map of fairly reliable data has been converted into an aesthetically more pleasing one, with no holes and apparently believable content. Validation is also problematic because verification is likely to be based on original polygon data rather than by commissioning new survey data, which is time-consuming and expensive. Thus, any variation generated within the polygon (a primary objective for the prediction in the first place) will be seen as a deviation from known data and may, therefore, be considered erroneous. On the other hand, pixel-by-pixel comparisons are equally invidious and error-prone, as the predictions used are statistically based and designed to be interpreted en masse rather than individually. This suggests that a high-resolution prediction can be effectively validated only when re-compiled to administrative-level summaries.

To minimize inconsistencies between original records and summed predicted values, two sets of standardized outputs have been produced in addition to the previously described 'raw' predictions.

## STANDARDIZING PREDICTED DISTRIBUTIONS

The numerical outputs of distribution modelling generally had mean values per polygon similar to those of the training data, but rarely matched exactly because regression analysis tends to smooth the peaks and troughs. In addition to the standardization imposed by the suitability masking, the following standardization procedures were adopted.

- model predictions for small polygons – defined as less than 1 000 km$^2$ – were replaced by suitability-corrected training data;
- model distributions were corrected so that totals calculated for training polygons matched the input training data, referred to as 'totals-corrected' distributions; and
- model values were adjusted so that calculated national totals matched the FAOSTAT country populations for year 2000, the so-called 'year 2000-corrected' distributions.

These corrections involved calculating a ratio between predicted and training data values for each polygon of observed (training) data and then applying the inverse of that ratio to the predicted data densities. The exception was where training data were absent, in which case predicted values were left unchanged.

Of the three routinely produced versions of livestock distribution based on suitability-corrected observed data, suitability- and totals-corrected, and suitability- and year 2000-corrected, the suitability- and totals-corrected version is the preferred output and is the version presented in the next section.

## PREDICTOR VARIABLES

A wide range of parameters has been incorporated in the analysis and modelling procedures, including ecoclimatic data, topography, human population data, cartographic data and data on protected areas.

### Satellite imagery

The livestock distribution modelling used the following satellite-derived measures of land-surface and atmospheric characteristics:

- NDVI from the AVHRR; a widely accepted measure of vegetation cover (Green and Hay, 2002; Hay, 2000; Hay *et al.*, 2006). Data were provided by the Pathfinder Program, initially supplied by the United States National Aeronautics and Space Administration's Global Inventory Monitoring and Modelling Systems group;
- a measure of land surface temperature derived by the TALA research group from thermal channels 4 and 5 of the AVHRR using the Price split window technique (Green and Hay, 2002; Hay, 2000; Hay *et al.*, 2006; Hay and Lennon, 1999; Price, 1984);
- a measure of air temperature (Tair), also derived from AVHRR channels (Goetz *et al.*, 2000);
- a measure of middle infrared reflectance, allied to temperature but less susceptible to atmospheric interference, derived from channel 3 of the AVHRR data (Hay, 2000);
- a measure of vapour pressure deficit derived from AVHRR channels 4 and 5 and ancillary processing (Green and Hay, 2002; Hay, 2000; Hay *et al.*, 2006); and
- a surrogate for rainfall – 'cold cloud duration' – derived from Meteosat remotely sensed data, provided by the FAO Artemis data archives (Hay, 2000).

All satellite-derived data were available as a series of decadal (ten-day) composite images, the AVHRR data covering an 18-year period from 1982 to 2000 and the Meteosat data covering a 29-year period from 1961 to 1990. Each series was subjected to temporal Fourier processing (named after the French mathematician, Joseph Fourier), re-sampled to 0.05-degree resolution (approximately 5 km$^2$ at the equator) and re-projected to the latitude/longitude system (geographic, or Plate Carrée projection). The Fourier processing of satellite data, described in detail in Rogers and Williams, 1994; Rogers *et al.*, 1996; Rogers, 1997; and Rogers, 2000, is quite central to the modelling process since it reveals the all-important seasonal characteristics of the environment. Each multi-temporal series is reduced to seven separate data layers: the mean, and the phases and amplitudes of the annual, bi-annual and tri-annual cycles of change. These are supplemented by three additional variables: the

**TABLE 5.1** GENERIC LIST OF VARIABLES USED IN LIVESTOCK DISTRIBUTION MODELLING

| Generic type | Variables |
|---|---|
| Locational | Longitude, latitude |
| Anthropogenic | Distance to roads [1] |
| | Distance to city lights [1] |
| Demographic | Human population [2] |
| Topographic | Elevation [3] |
| Land cover | Normalized difference vegetation index [4, 5, 6] |
| Temperature | Land surface temperature [4, 5, 6, 7, 8] |
| | Air temperature [9] |
| | Middle-infrared [5] |
| Water and moisture | Vapour pressure deficit [4, 5, 6] |
| | Distance to rivers [10] |
| | Cold cloud duration [5, 11] |
| | Potential evapotranspiration [11] |
| General climatic | Modelled length of growing period [12] |
| Other | Tsetse distributions (for Africa) [13] |

[1] Derived from layers in the LandScan archive, produced and distributed by Oak Ridge National Laboratories (ORNL) (http://www.ornl.gov/sci/gist/projects/LandScan).
[2] Taken from CEISIN's Gridded Population of the World (GPW) version 2 dataset (http://sedac.ciesin.columbia.edu/gpw).
[3] Global GTOPO30 1km resolution elevation surface, produced by the Global Land Information System (GLIS) of the United States Geological Survey, Earth Resources Observation Systems (USGS, EROS) data centre (http://edc.usgs.gov/products/elevation/gtopo30/gtopo30.html).
[4] Green and Hay, 2002.
[5] Hay, 2000.
[6] Hay *et al.*, 2006.
[7] Hay and Lennon, 1999.
[8] Price, 1984.
[9] Goetz *et al.*, 2000.
[10] Derived from the USGS EROS data centre HYDRO 1k data archive (http://edc.usgs.gov/products/elevation/gtopo30/hydro/index.html).
[11] Mean, minimum and maximum decadal estimates of 'cold cloud duration' were derived from METEOSAT remotely-sensed data (1961-90), obtained from FAO's Artemis data archives.
[12] Fischer *et al.*, 2002 (http://www.fao.org/waicent/faoinfo/agricult/agl/agll/gaez/index.htm).
[13] Tsetse distributions used were those developed for the Programme Against African Trypanosomiasis (PAAT) Information system (http://www.fao.org/ag/againfo/programmes/en/paat/infosys.html).

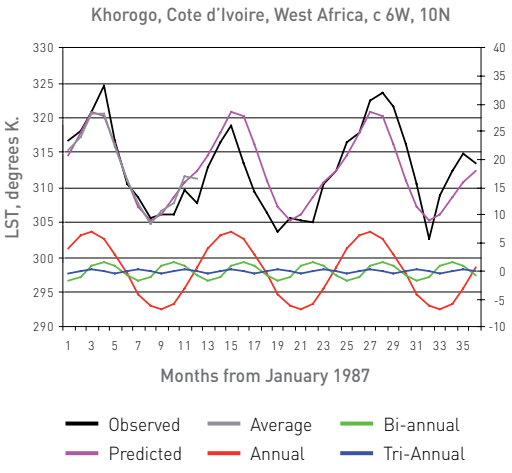
*Source:* Robinson *et al.*, in press.

maximum, the minimum and the variance of the satellite-derived geophysical variables.

The temporal Fourier processing of multi-temporal data is illustrated in Figures 5.5 (land surface temperature) and 5.6 (NDVI), both taken from FAO, 2006b. In each case, three years of monthly AVHRR data are shown as the black lines (the additional grey line in year 1 is the three-year average). The annual, bi-annual and tri-annual Fourier cycles are shown in red, green and blue, respectively (notice the second, zero-centred scale for these on the upper graph, right-hand axis), and their re-combined sum is shown as the violet line super-imposed on the raw data. These figures illustrate how the Fourier decomposition manages to capture subtle details of the seasonal cycle in both variables.

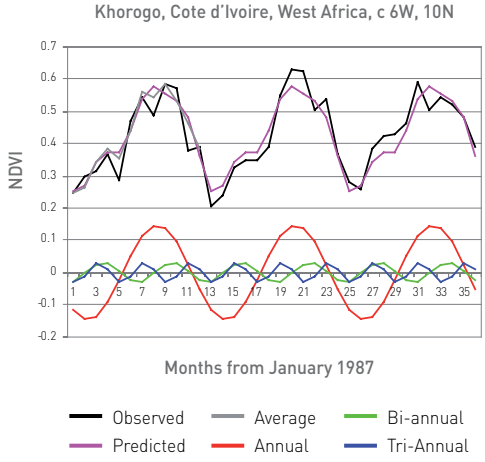
The Fourier variables were calculated and turned into GIS image data layers, together with the maximum and minimum values and variances of each original signal. Collectively, these numerical indictors of the level (mean, minimum, maximum), timing (phase), seasonality (amplitude) and variability (variance) of each satellite-derived

**5.5** FOURIER-PROCESSED LAND SURFACE TEMPERATURE TIME SERIES FROM A SINGLE POINT IN NORTHERN CÔTE D'IVOIRE




*Source:* FAO, 2006b.

**5.6** FOURIER-PROCESSED NDVI TIME SERIES FROM A SINGLE POINT IN NORTHERN CÔTE D'IVOIRE




*Source:* FAO, 2006b.

environmental variable give a unique 'fingerprint' of habitat type; they provide a link between the satellite signal and the biological processes that are, in one way or another, linked to the suitability of the environment to support livestock. A further advantage of the Fourier processing is that it reduces the vast number of individual decadal images to a manageable and relatively independent set of variables, more amenable to statistical analysis and interpretation.

The power of these Fourier-processed data to distinguish habitat types is illustrated in Figure 5.7, taken from Rogers and Robinson, 2004, in which three of the Fourier variables for the NDVI images for Africa are combined as a false colour composite: the average value (or 'zero-order' component) is displayed in red; the phase of the first-order component is displayed in green; and the amplitude of the first-order component is displayed in blue.

### Other eco-climatic and land-related data

Elevation data were obtained from the USGS EROS data centre's GTOPO30 1 km resolution DEM for Africa[31]. A series of land-use variables were extracted from the LandScan data set[32], including slope and vegetation cover. In addition, rivers were taken from the USGS EROS data centre's HYDRO 1k data archive[33]. Larger rivers were identified according to their flow accumulation characteristics, from which a distance-to-rivers image was prepared.

Potential evapotranspiration and annual rainfall data were taken from the FAO/IIASA data archive (Fischer *et al.*, 2002)[34] and re-sampled to a 0.05-degree resolution.

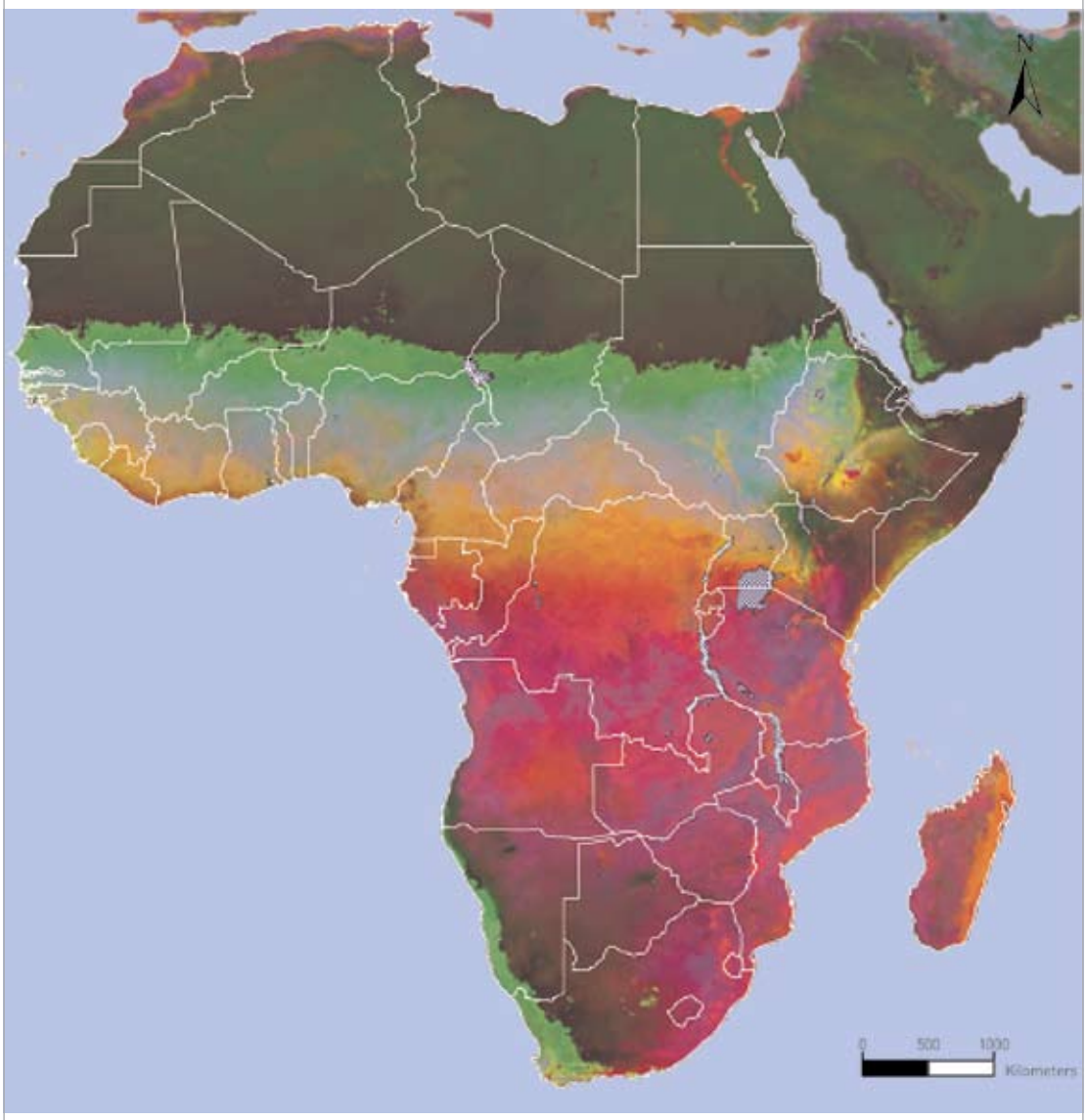
The LGP was modelled separately for each continent, using regression techniques illustrated earlier in this section and the FAO/IIASA archive values as training data.

### Human population data

As the GLW project has evolved, so also have the sources of human population data used in the modelling. Early on, for Africa and Asia, human

[31] http://edc.usgs.gov/products/elevation/gtopo30/gtopo30.html
[32] http://www.ornl.gov/sci/gist/projects/LandScan
[33] http://edc.usgs.gov/products/elevation/gtopo30/hydro/index.html
[34] http://www.fao.org/waicent/faoinfo/agricult/agl/agll/gaez/index.htm

**5.7** FALSE COLOUR COMPOSITE OF FOURIER-PROCESSED NDVI VARIABLES FOR AFRICA




*Source:* Rogers and Robinson, 2004.

population data were derived from three sources: (i) estimates collated by the FAO Agriculture Land and Water Division at five-minute resolution; (ii) data, again at five-minute resolution, provided by the Centre for International Earth Science Information Network (CIESIN), derived from data collated by the National Centre for Geographic Information and Analysis[35]; and (iii) data for the Horn of Africa came from the Intergovernmental Authority on Drought and Development – now known as the Intergovernmental Authority on Development (Wint *et al.*, 1997). Pixel values from these sources were averaged.

[35] http://www.ncgia.ucsb.edu/pubs/gdp/pop.html

More recently efforts have been made to compile global human population data: first, the LandScan project[36] and, more recently, CIESIN's Gridded Population of the World[37], which is now in its third version and includes the Global Rural-Urban Mapping Project datasets. In the more recent analyses, the project has moved towards these more consistent datasets.

Other data related to human population distributions and proximity to night-time lights and roads were generated from layers available in the Columbia University LandScan archive[38].

[36] http://www.ornl.gov/sci/gist/projects/LandScan
[37] http://sedac.ciesin.columbia.edu/gpw
[38] http://www.ornl.gov/sci/gist/projects/LandScan