PATH and Tableau Foundation #VisualizeNoMalaria Project:

Development of an automated predictive and forecasting malaria cases capability

Author: Allan Walker

Contributing Peers: Anya A'Hearn, Chris DeMartini, Douglas Morris, Joe Mako, Philip Riggs

Sponsor, Reviewer: Jeff Bernson

A note from PATH

Malaria is a disease that does not discriminate. Male or female, young or old, rich or poor—the malaria parasite will happily take up residence in the bloodstream of any human being without consent and bring pain, suffering, and possible death to its host.

Efforts to wipe out malaria have long focused on the mosquitos responsible for transmitting the disease, but the parasite only spends one-half of one percent of its life anywhere outside the human bloodstream. But despite the easily treatable nature of the disease, only a fraction of people—approximately 20%—carrying the malaria parasite in their bloodstream are symptomatic and seek out treatment. If we have any hope of eliminating the disease—particularly in Sub-Saharan Africa, home to 88 percent of global malaria cases in 2015—we must find new ways to identify, treat, and monitor the estimated four in five infected people who are unknowingly carrying the disease.

In Zambia alone, it's estimated that nearly 2,300 people die from malaria each year—the majority of which are children. For over a decade, Zambia's Ministry of Health has aggressively fought the disease—and made tremendous gains—with steadfast political support across the country's political spectrum. The Zambian government also has the backing of united partners including the The Global Fund to Fight AIDS, Tuberculosis and Malaria, the U.S. President's Malaria Initiative, the World Bank, as well as the PATH-led Malaria Control and Elimination Partnership in Africa (MACEPA), which has supported the Ministry's efforts with lifesaving tools and training.

In that time, Zambia's National Malaria Control Program, in partnership with PATH, tested a combination of interventions to protect people from malaria. Insecticide-treated bed nets, indoor residual spraying, mass drug administration, and now sophisticated data analytics systems have all been deployed with great success. Between 2012 and 2015, Southern Province has seen a 93 percent reduction in malaria in children, and a 97 percent reduction in malaria-related deaths for people of all ages.

However, as impressive as those stats are, hurdles remain before the fight will be won. Tracking down the remaining parasites—those in the bloodstreams of asymptomatic people who continue to live their lives in their communities and travel throughout their country—remains an incredible challenge that calls for continued innovation.

The sophisticated predictive modeling work described in this paper may prove to be another one of those critical interventions. Up until now, finding asymptomatic carriers relied on figuring out where the symptomatic carriers had been in the recent past. After a period of testing and improvement, the hope is that the models could show where the asymptomatic and symptomatic carriers could be in the near future. This information would allow officials and health workers at every level to get ahead of the disease and deploy resources accordingly.

Plenty of research, testing, and refinement will go into the process before we'll know the true value of these models in fighting the disease. But with so many lives on the line, we'll continue exploring opportunities to give everyone in the country—male or female, young or old, rich or poor—an opportunity to live a malaria-free life.

To learn more about the disease and the work in Zambia to eliminate it, please visit visualizenomalaria.org.

- Jeff Bernson, *Director, Results Management, Measurement & Learning at PATH*

Table of Contents

Capability Statement

This capability was developed to initially support District Health Management Teams in Zambia (and potentially scale outwards to other countries such as Senegal and Ethiopia) by providing an automated, timely, and robust forecast of malaria cases within their catchment area via an interactive Tableau dashboard.

Scope

The scope of this project is currently limited to a Proof of Concept, albeit delivery of an initial operating capability post-review and -user acceptance testing.

The capability currently considers records of active health facilities, reported weekly, in the Southern Province of Zambia.

Target Audience

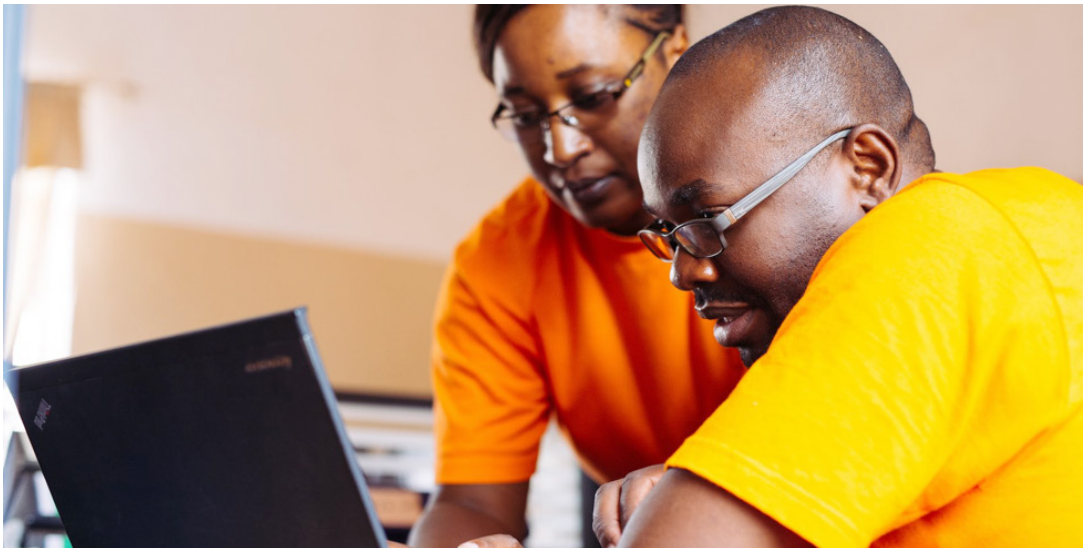


Figure 1 "Marie-Reine Rutagwera, malaria surveillance specialist for MACEPA, Monde Mathews, environmental health officer for Gwembe district, reviewing malaria surveillance data on a laptop." Photograph: Gabe Bienczycki/PATH/Gabe Bienczycki

This capability is designed to help Health Facility decision-makers plan and manage resources.

However, it is not intended to replace local knowledge. Regardless of how accurate a prediction may be, or how robust a forecast model is, the author notes that this capability should only act as an enhancement to the sterling work that is being done in the field, and there is no substitute for practical experience.

Product

Interactive Situation Awareness Dashboard

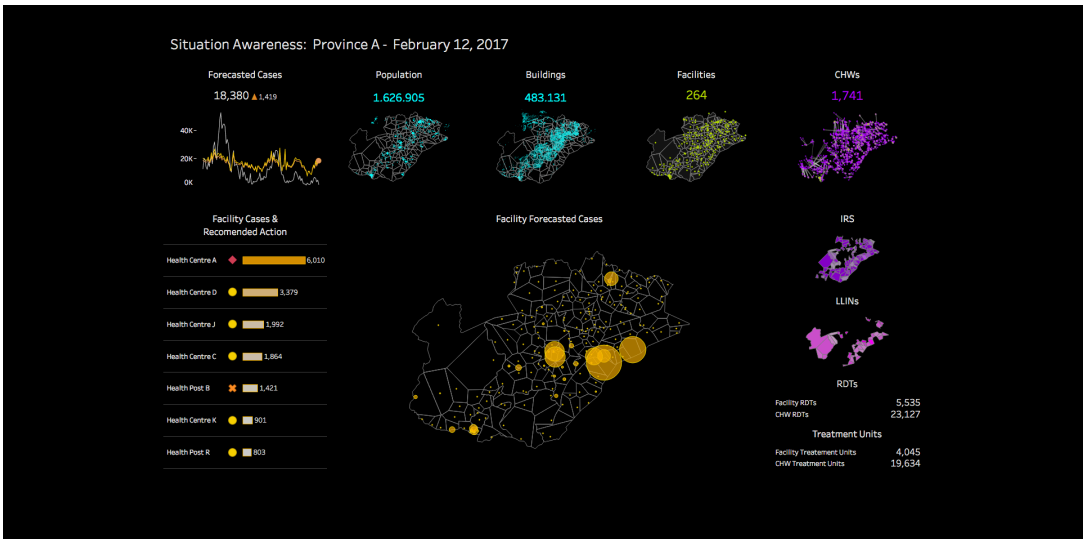


Figure 2 Zambia Southern Province Situation Awareness using Tableau Software, courtesy Anya A'Hearn

Description

Situational Awareness (SA) is the ability to identify, process, and comprehend the critical elements of information about what is happening to the team with regards to the mission. (Guard, n.d.).

The intent of these dashboards is to highlight recommended insights for users to consume and present recommended actions to enhance decision making.

A further intent is to enable users to drill from Province to District to an individual Health facility, enabling SA at each geographic hierarchical level.

Figure 3 Generic Facility Situation Awareness Dashboard using Tableau Software, courtesy Anya A'Hearn

Automated Alteryx Workflow

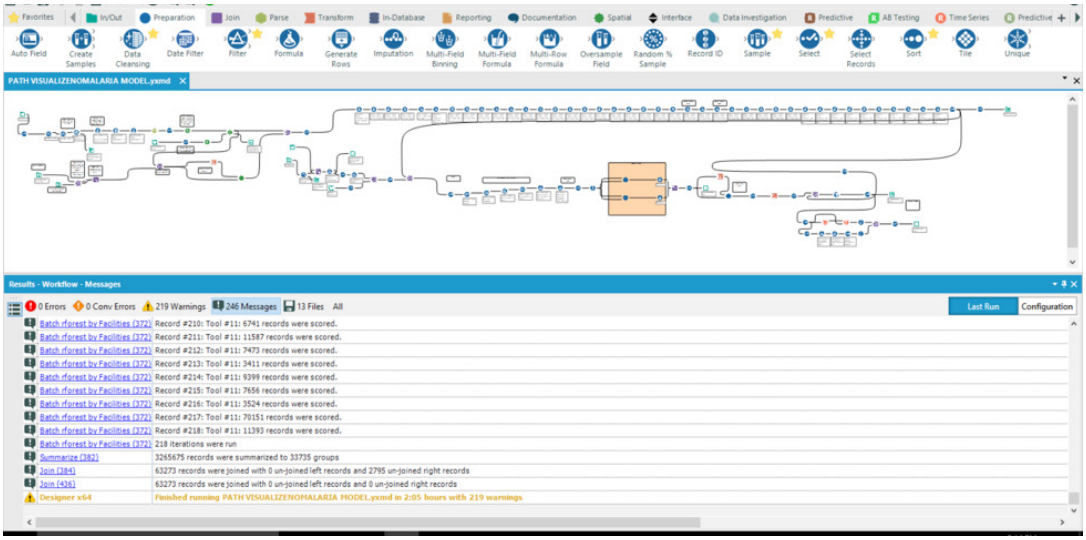


Figure 4 Automated Alteryx Workflow

Description

An Alteryx workflow was developed to enable automated prediction and forecasting of malaria cases in the Southern Province of Zambia. It can be batch executed on a weekly cadence.

In-memory Database

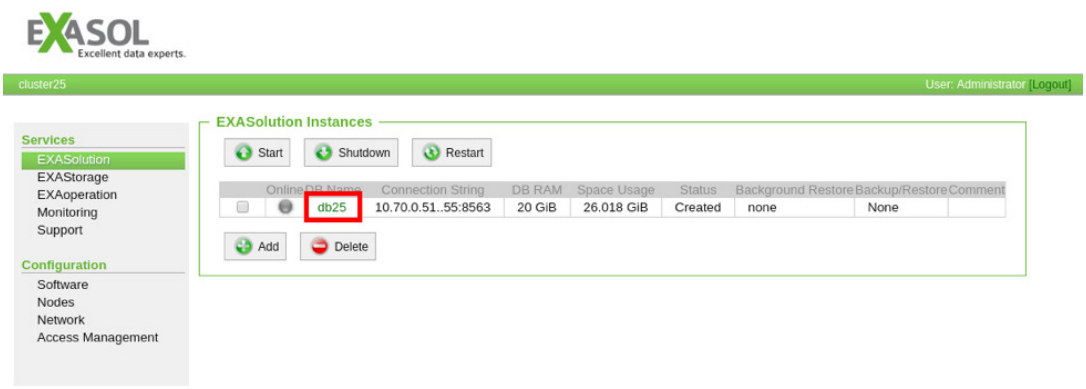


Figure 5 ExaSolution Management Panel

An Exasol database was developed to store the input variables and the output data. It connects both to Alteryx as an in-memory data source (for the workflow described above) and as a live data source to Tableau (for the dashboards as hitherto mentioned).

Architecture

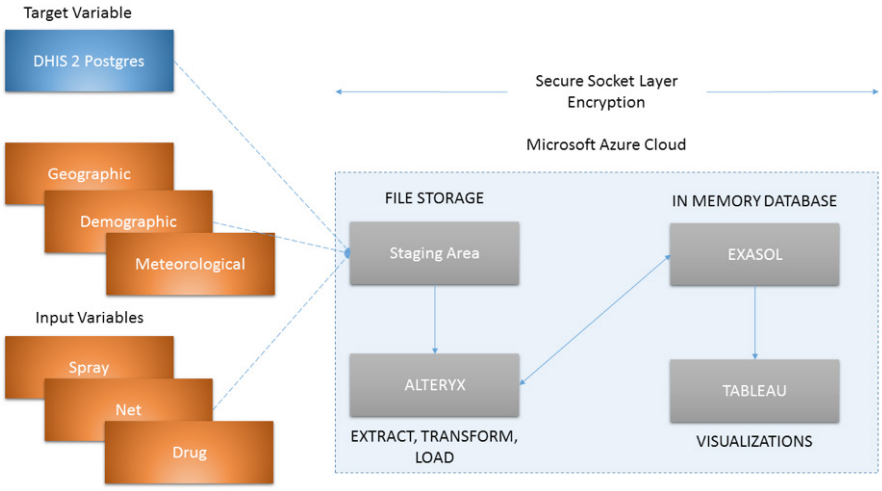


Figure 6 Proposed Architecture/Technology Stack

Data

The data used for this capability is detailed in the "Input Variables" section of this document: extensive and detailed analysis of geographies, including topology, hydrology, land cover, and population densities and migration. Spray, Net, and Drug administration data is intended to improve the fidelity of the capability, though this has not been completed at the time of publication. Likewise, it is further envisioned to add additional input variables such as meteorology data, vegetation indices, and population mobility, depending on availability and suitability.

Cloud

This capability will leverage the Microsoft Azure Cloud for data storage, extract-transform-load, and visualization. This will minimize latency between the servers, and provide inherent security with Secure Socket Layer (SSL) certification.

Technologies and Software

Tableau

Tableau Software is a software company headquartered in Seattle, Washington, which produces interactive data visualization and analytics products. Tableau's mission statement: help people see and understand their data.

QGIS

QGIS (previously known as Quantum GIS) is a cross-platform, free, and open-source desktop geographic information system (GIS) application that supports viewing, editing, and analysis of geospatial data.

Alteryx

Alteryx is a software company headquartered in Irvine, California, with a development center in Broomfield, Colorado. The company's products are used for data blending and advanced data analytics. Alteryx has a stated goal of enabling advanced analytics to be performed by non-specialists.

EXASOL

EXASOL is an analytic database management software company. Its product is called EXASolution, an in-memory, column-oriented, relational database management system.

Mapbox

Mapbox is the location data platform for mobile and web applications. The tool provides building blocks to add location features like maps, search, and navigation into any experience you create.

Mapbox is changing the way people move around cities and explore our world. Mapbox's apps reach more than 300 million people each month.

DHIS 2

DHIS 2 is the preferred health management information system in 47 countries and 23 organizations across four continents. DHIS 2 helps governments and health organizations to manage their operations more effectively, monitor processes and improve communication.

About #VisualizeNoMalaria

Project

A partnership with PATH, the Tableau Foundation, and you can help make malaria history

PATH is an international health organization driving transformative innovation to save lives. Working in partnership with national governments, PATH is leading the way toward a malaria-free world by focusing on new vaccines, treatments, diagnostics, and approaches.

Tableau has partnered with PATH to provide vital software, training, and funding in support of Zambia's goal of eliminating malaria by 2020. Our partnership will empower frontline health workers with the critical tools to track and treat malaria cases to help eliminate this deadly disease.

Together, we are improving data accuracy and making critical data-informed decisions about how and where to tackle outbreaks. We are also building the skills of district and facility health teams to combat the disease at the community level. If successful, this model could be scaled globally to end malaria for good.

Team

Republic of Zambia Ministry of Health

The Government of the Republic of Zambia through its ongoing health sector reforms aims to improve health outcomes.

PATH

PATH is the leader in global health innovation. An international nonprofit organization, we save lives and improve health, especially among women and children. We accelerate innovation across five platforms—vaccines, drugs, diagnostics, devices, and system and service innovations—that harness our entrepreneurial insight, scientific and public health expertise, and passion for health equity. By mobilizing partners around the world, we take innovation to scale, working alongside countries primarily in Africa and Asia to tackle their greatest health needs. Together, we deliver measurable results that disrupt the cycle of poor health.

Tableau Foundation

The Tableau Foundation is an initiative led by the employees of Tableau Software that encourages the use of facts and analytical reasoning to solve the world's problems. Tableau Foundation grants combine Tableau's most valuable resources—its people, its products, and its community—with financial support to nonprofits that are using data to reshape communities around the globe.

The VisualizeNoMalaria project is a landmark case of the Tableau Foundation's culture of collaboration. Software companies, consulting partners—including several Zen Masters, Tableau's most accomplished users—and individual volunteers from across the globe have contributed their technology, talents, and time to the campaign.

Alteryx for Good

The Alteryx for Good program provides students and educators with a free Alteryx Designer license to help foster learning and further classroom teaching. Likewise, small non-profits and government agencies can receive licenses to aid in achieving their respective missions and goals. Our customers, partners, and employees are enabling education of next-gen analysts, influencing social change through analytics, and inspiring people to be a part of a greater purpose leveraging our community of analytic volunteers.

DataBlick

A boutique consulting firm that transforms your data into actionable art. DataBlick wrangles data, wraps people's brains around complex math, and makes the result stunning.

EXASOL

The company develops the world's fastest database for analytics and data warehousing, EXASolution, and offers first-class know-how and expertise in data insight and analytics.

Mapbox Humanitarian

Tools that let you build fast with massive data. Mapbox Humanitarian is committed to supporting humanitarian responders with data, satellite imagery, services for OpenStreetMap, and free Mapbox accounts.

About the author

Allan Walker is a volunteer working in his spare time on the PATH #VisualizeNoMalaria project.

Credits

Thank you to the authors of previous research, whose work was the inspiration for this capability. The first document to credit, "Topographic models for predicting malaria vector breeding habitats: potential tools for vector control managers" (Nmor JC, 2013) is particularly significant, as the results and conclusions suggested that acquiring suitable topographic raster data with further processing could "sufficiently accurate to predict vector habitats".

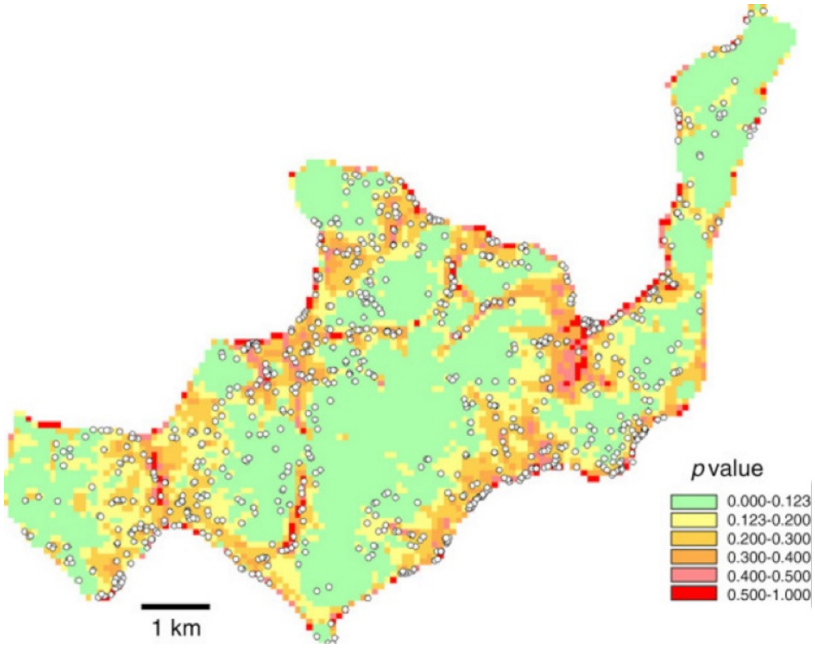


Figure 7 SRTM model: "The likelihood of the presence of breeding sites in Rusinga based on logistic regression modeling" © 2013 Nmor et al; licensee BioMed Central Ltd

The second article, "Local topographic wetness indices predict household malaria risk better than land-use and land-cover in the western Kenya highlands" (Justin M Cohen, 2010) concluded:

Topographic wetness values in this region of highly varied terrain more accurately predicted houses at greater risk of malaria than did consideration of land-cover/land-use characteristics. As such, those planning control or local elimination strategies in similar highland regions may use topographic and geographic characteristics to effectively identify high-receptivity regions that may require enhanced vigilance.

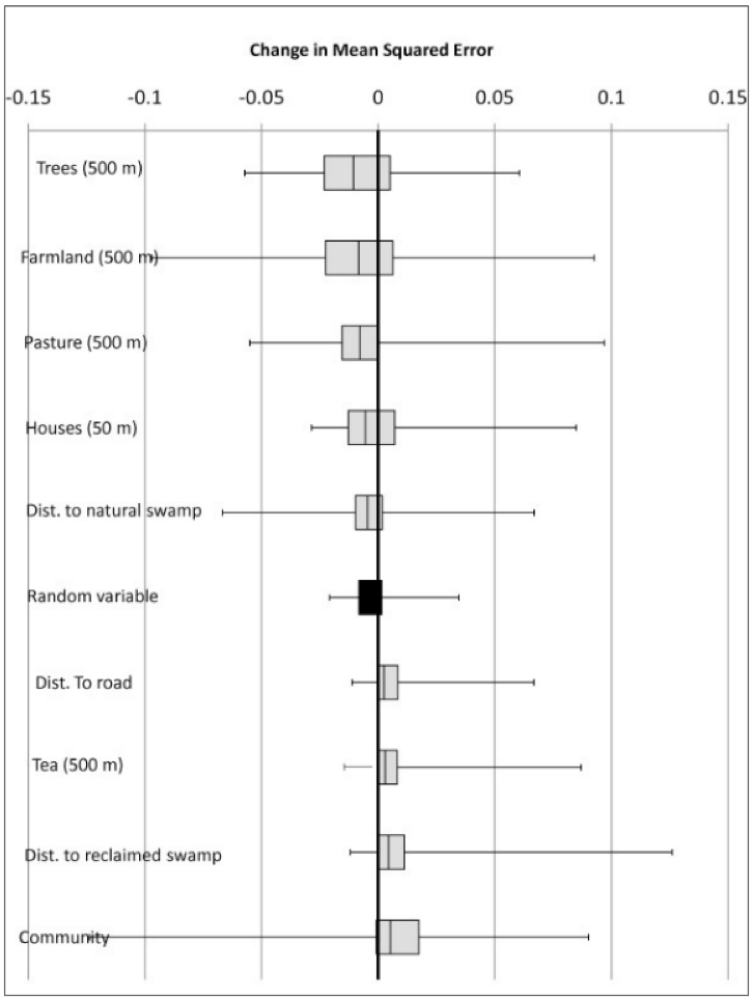


Figure 8 "Change in mean-squared-error resulting from the addition of land-cover/land-use variables to the best-fitting topographic predictive model for each random subset of houses" ©2010 Cohen et al; licensee BioMed Central Ltd.

Nota Bene

The author notes that there is extensive and ongoing research into predicting vector-borne diseases and as such recognizes and commends all research. The author does not make any claim of intellectual property, and furthermore, states no competing interest.

The author makes no claim to be a trained data scientist, or an accomplished mathematician. There is therefore recognizable room for improvement in the capability, exempli gratia:

1. The Random Forest is set using default configurations and parameters, and
2. The ARIMA Forecast is set using default configuration and parameters which therefore offers opportunities for ongoing improvement and fidelity of results.

Input Variables

Elevation

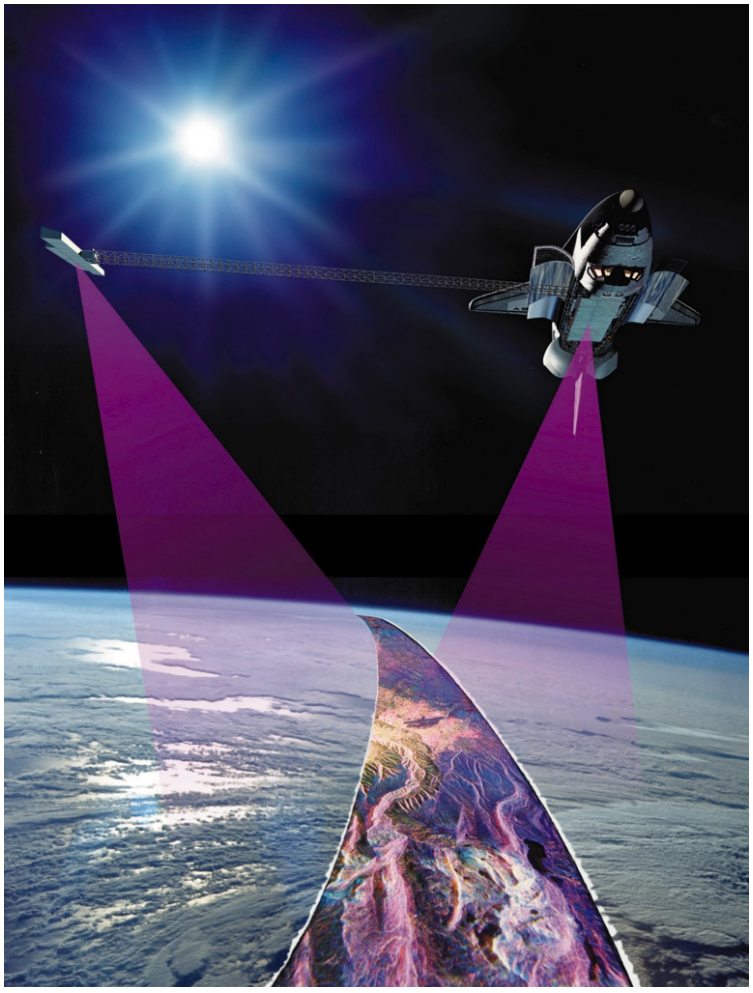


Figure 9 "Observing and Mapping the Earth" courtesy NASA/JPL-Caltech

Definition

The NASA Shuttle Radar Topographic Mission (STRM) derived digital elevation models (DEM) in raster format at a resolution of 90 meters (at the equator) are provided by the Consortium of International Agricultural Research – Consortium for Spatial Information (CGIAR-CSI) in 5 degree x 5 degree tiles. (SRTM 90m Digital Elevation Database v4.1, n.d.)

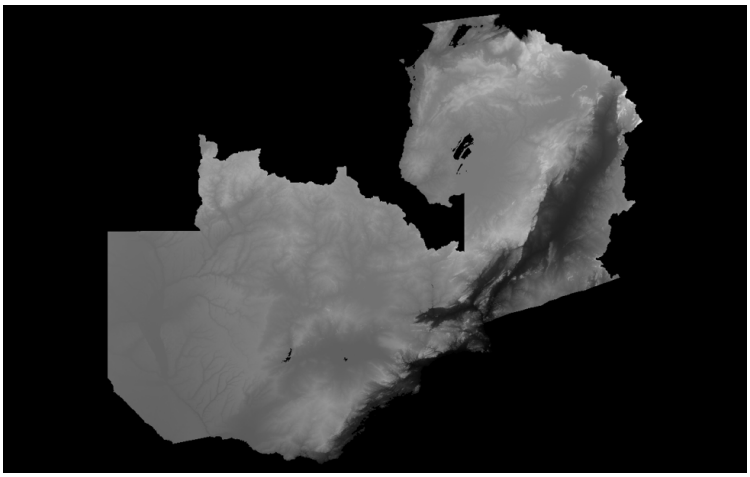


Figure 10 Elevation of Zambia, raster GeoTIFF rendered using QGIS

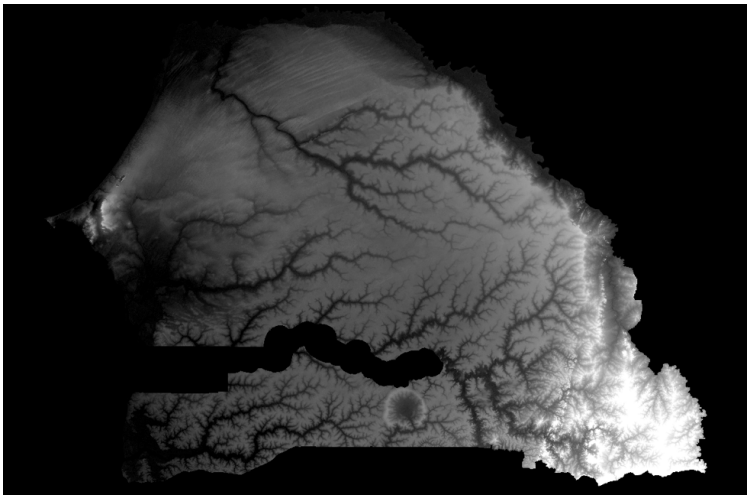


Figure 11 Elevation of Senegal, raster GeoTIFF rendered using QGIS

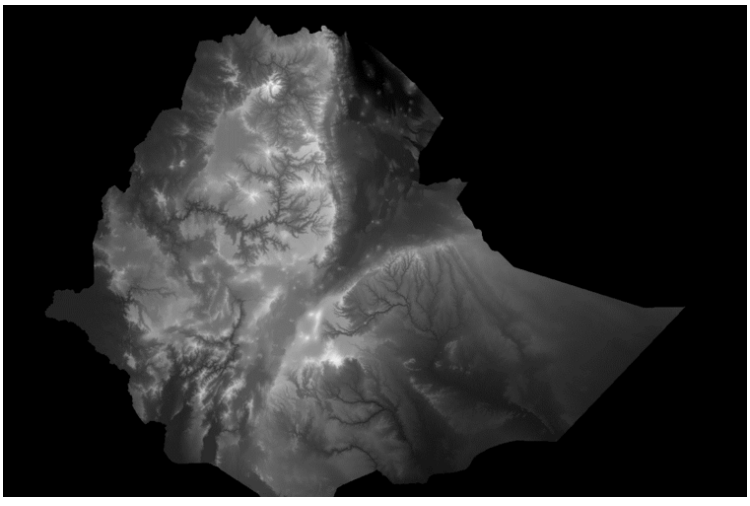


Figure 12 Elevation of Ethiopia, raster GeoTIFF rendered using QGIS

Method

STRM (Shuttle Topographical Radar Mission) GeoTIFF raster files were obtained from the CGIAR-CSI (Consortium of International Agricultural Research – Consortium for Spatial Information) repository. Using GDAL (Geospatial Data Abstraction Library) commands in QGIS, the GeoTIFF raster files were merged and built into a virtual raster (VRT). These VRT files were then warped (re-projected) from WGS84 EPSG:4326 into WGS84 UTM Zone 36S (for Zambia), WGS84 UTM Zone 28N (for Senegal) and WGS84 UTM Zone 37N (for Ethiopia). The VRT files were then translated back to GeoTIFF format.

Vector polygon shapefiles of Zambia and Senegal at the country administration level were obtained from the FAO (Food and Agriculture Organization of the United Nations) repository and re-projected from WGS84 EPSG:4326 to WGS84 UTM Zone 36S (for Zambia), WGS84 UTM Zone 28N (for Senegal) and WGS84 Zone 37N (for Ethiopia).

To obtain the outline of the countries, the three GeoTIFF files were then clipped by mask layer, cropping the extent of the target database to the extent of the cutline, while retaining the resolution of the output raster.

To obtain a vector array of X (Longitude in decimal degrees), Y (Latitude in decimal degrees), and Z (Elevation in meters), the polygonize GDAL command in QGIS was executed on the clipped GeoTIFF files. Using SAGA GIS (Python) command polygon centroids, point vector shapefiles were generated. The shapefiles were re-projected back from UTM Zones 36S (for Zambia), 28N (for Senegal), and 37N (for Ethiopia) to WGS 84 EPSG:4326 and exported to CSV (Comma Separated Value) files.

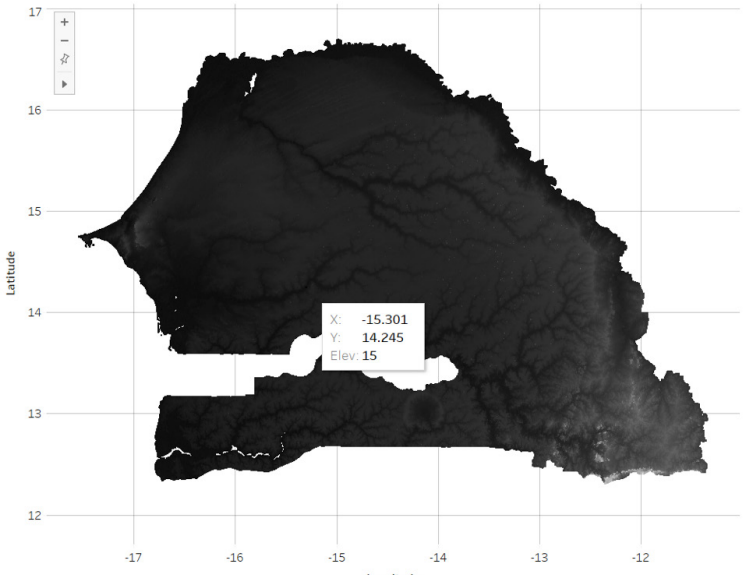


Figure 13 Elevation of Senegal, vector array rendered using Tableau Software

Slope and Catchment Area

Slope

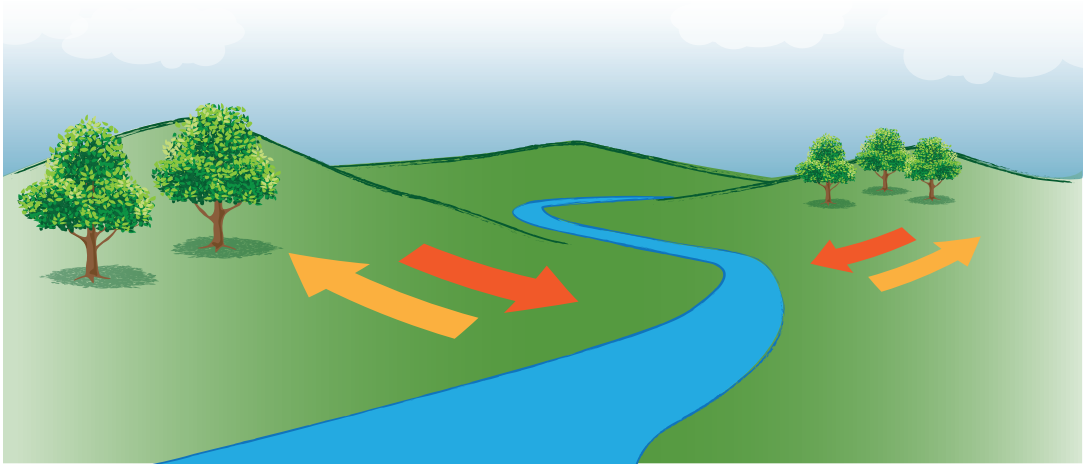


Figure 14 "A rising and a falling slope" courtesy Natural Resources Management and Environment Department, Food and Agriculture Organization

Definition

Slope S is defined as the magnitude of vector quantity with components equal to the partial derivatives of the surface in the x and y directions (gradient), or the tangent of the angle of the steepest slope of a plane tangential to the surface. (J. & F., 1997).

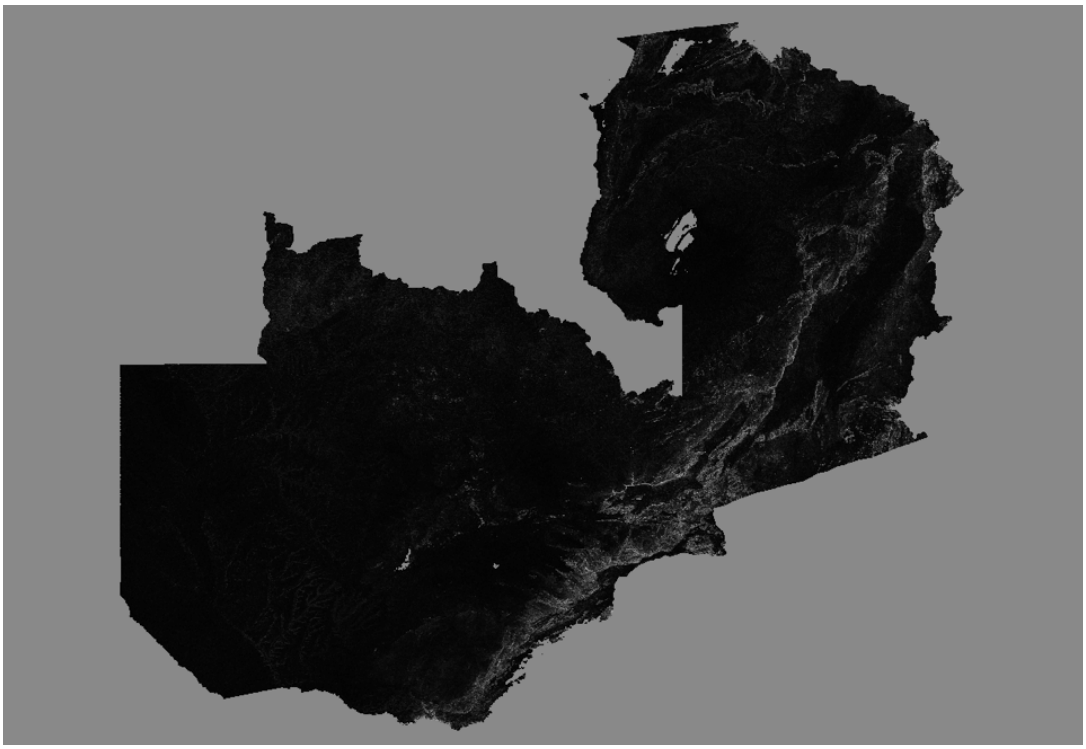


Figure 15 Slope of Zambia, raster GeoTIFF rendered using QGIS

Catchment Area

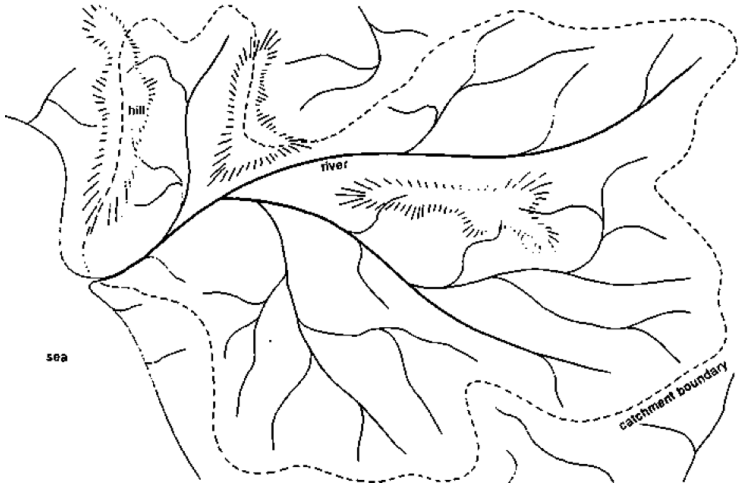


Figure 16 "Catchment Area of a River" courtesy Natural Resources Management and Environment Department, Food and Agriculture Organization

A drainage basin (watershed) is a portion of the Earth surface occupied by a main stream and its tributaries separated from adjacent basins by a drainage divide. (Wisconsin–Stevens)

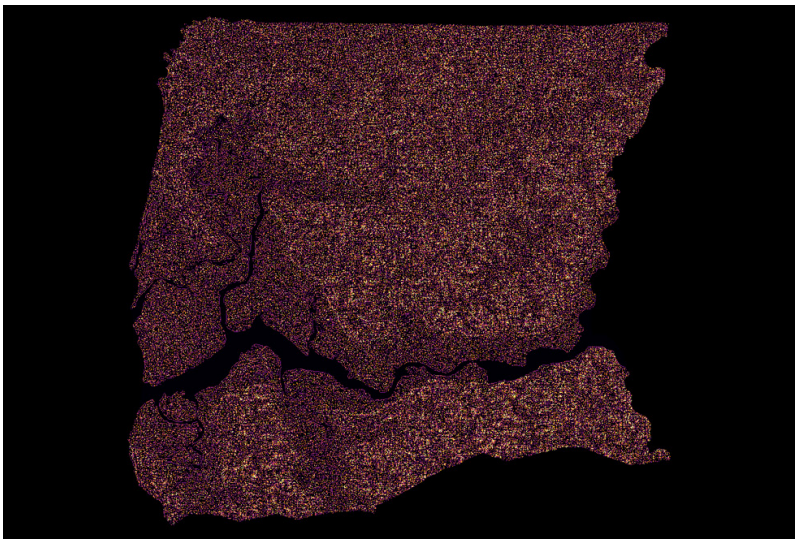


Figure 17 Catchment Area of Ziguinchor Region of Senegal, raster GeoTIFF rendered using QGIS

Method

To obtain Slope and Catchment Areas, the Elevation GeoTIFFs for Senegal and Zambia were initially processed with the SAGA GIS (Python) Fill Sinks (Wang & Liu) algorithm. This outputs three GeoTIFFs: Filled DEM (Digital Elevation Model), Flow Directions, and Watershed Basins.

The SAGA GIS Catchment Area (recursive) algorithm is then executed, using the Filled DEM as the Elevation parameter. Using the default options (choosing Deterministic 8 as the method), outputs of Catchment Slope and Catchment Area are chosen.

To obtain a vector array of X (Longitude in decimal degrees), Y (Latitude in decimal degrees), and Z (Slope), the polygonize GDAL command in QGIS was executed on multiple clipped GeoTIFF files. Due to the size of the vector shapefiles produced, Regional (Senegal & Ethiopia) and Provincial (Zambia) administration level masks were used.

Using SAGA GIS command polygon centroids, point vector shapefiles were generated. The shapefiles were re-projected back from UTM Zones 36S (for Zambia), 28N (for Senegal), and 37N (for Ethiopia) to WGS 84 EPSG:4326 and exported to CSV. An Alteryx workflow was developed to union the Regional and Provincial CSVs to produce a single CSV for each Country.

Topographic Wetness and Stream Power Indices

Topographic Wetness Index

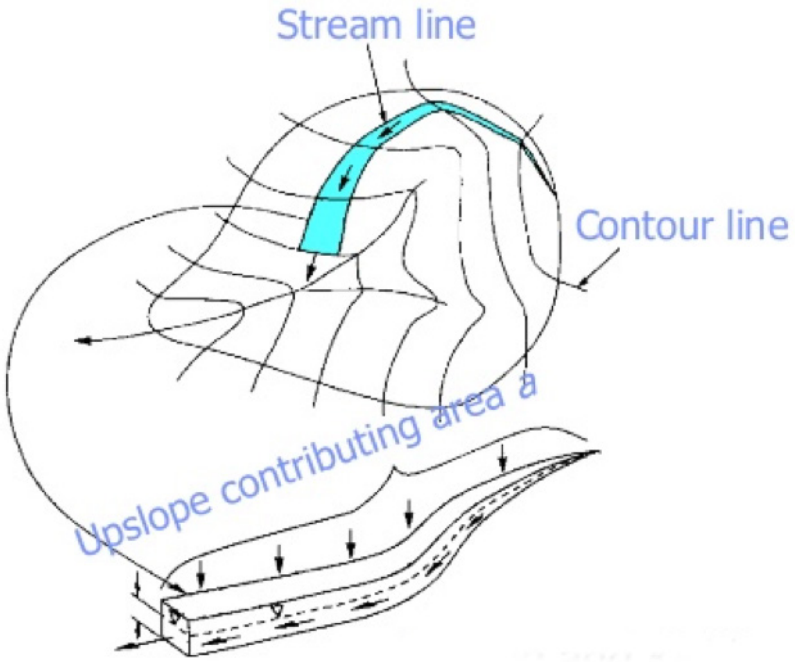


Figure 18 "The concept of contributing area per unit" reference Bevan and Kirkby, courtesy Salvatore Manfreda, Università degli Studi della Basilicata

Definition

The topographic wetness index (TWI, ln(a/tanß)), which combines local upslope contributing area and slope, is commonly used to quantify topographic control on hydrological processes. (R. Sørensen, 2006)

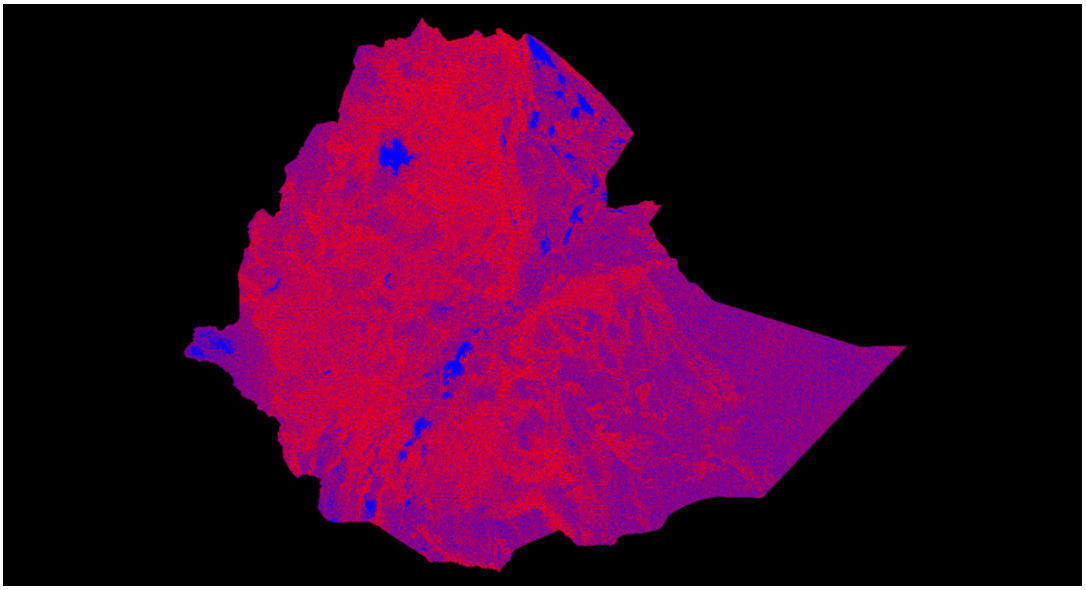


Figure 19 Topographic Wetness Index of Ethiopia, raster GeoTIFF rendered using QGIS

Stream Power Index

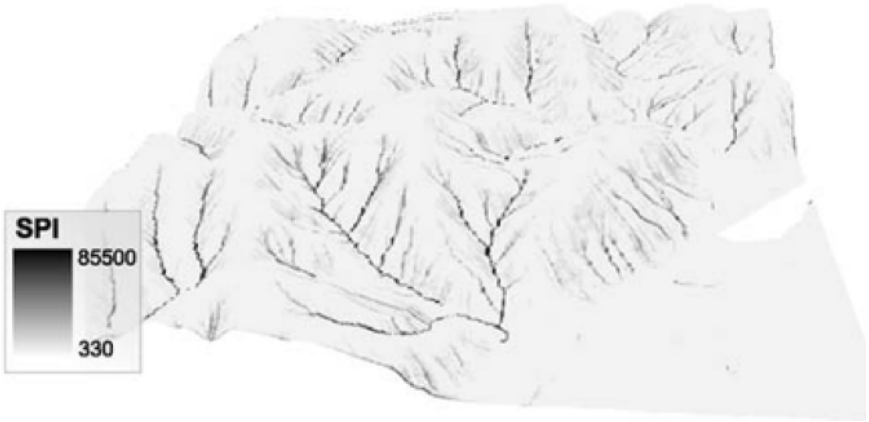


Figure 20 "Stream power index maps of the Baranja Hill Case study with a resolution of 10 m", reference Developments in Soil Science, courtesy Hannes Isaak Reuter (European Commission) & Andy Nelson (University of Twente)

Definition

The Stream Power Index (SPI) is a measure of the erosive power of flowing water. SPI is calculated based upon slope and contributing area. (Gallant, 2000)

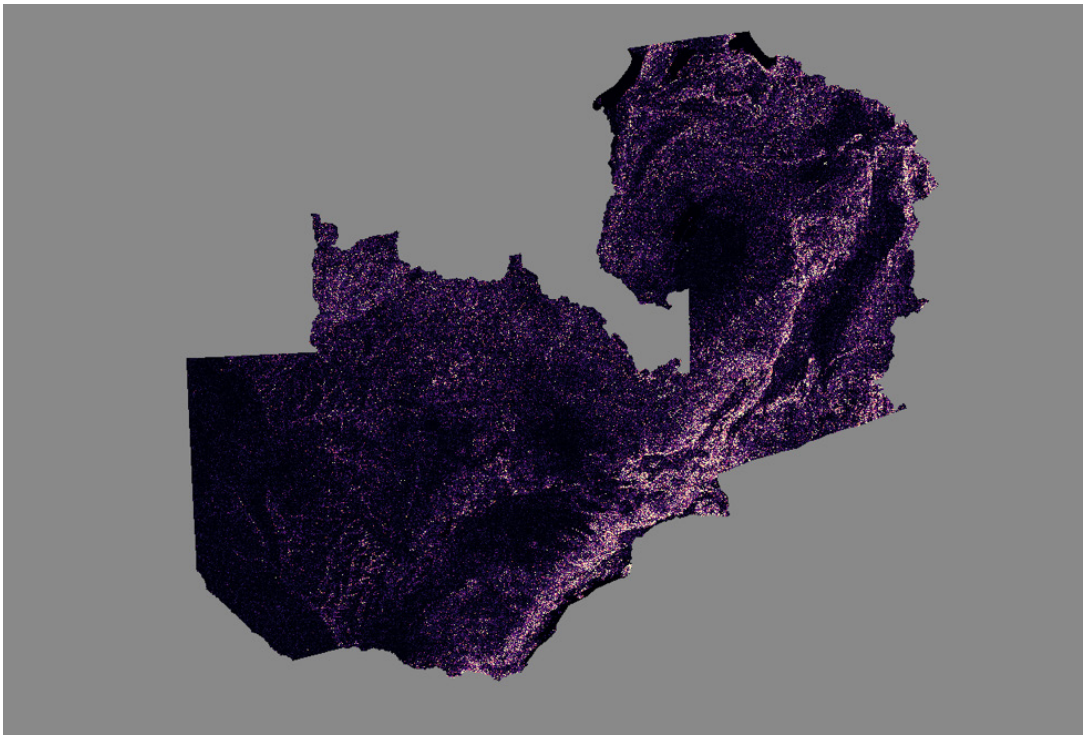


Figure 21 Stream Power Index of Zambia, raster GeoTIFF rendered using QGIS

Method

To obtain the Topographical Wetness Index (TWI) and Stream Power Indices (SPI) for Senegal, Ethiopia, and Zambia, SAGA GIS (Python) algorithms were executed using the Slope and Catchment Area GeoTIFFs as input parameters respectively.

Vector arrays for both TWI and SPI were generated using the procedure mentioned hitherto.

Channel Network and Strahler Order

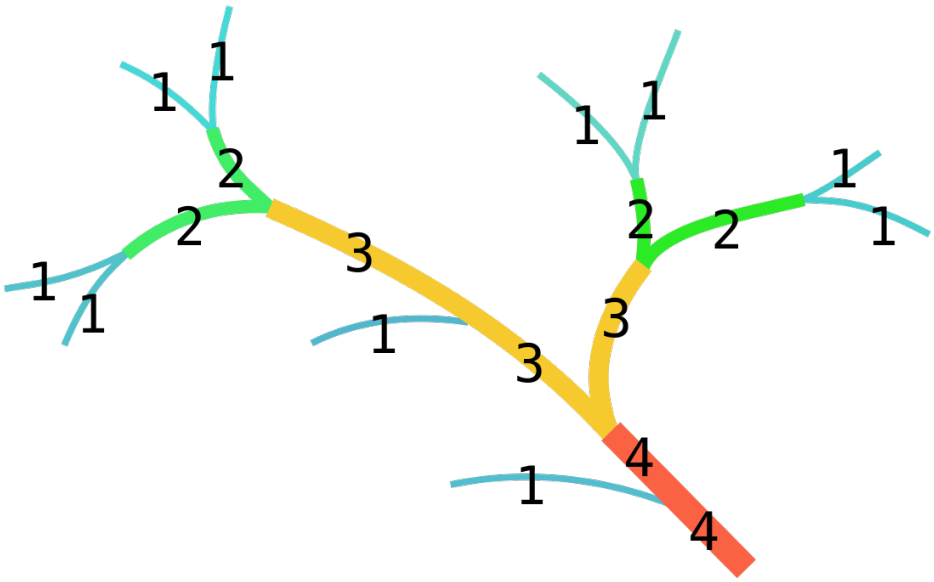


Figure 22 "Diagram showing the Strahler stream order" courtesy Wikipedia

Definition

A stream is classified as a body of water that flows across the Earth's surface via a current and is contained within a narrow channel and banks. Strahler outlined the order of streams to define the size of perennial (a stream with water its bed continuously throughout the year) and recurring (a stream with water in its bed only part of the year) streams. (STRAHLER, 1952)

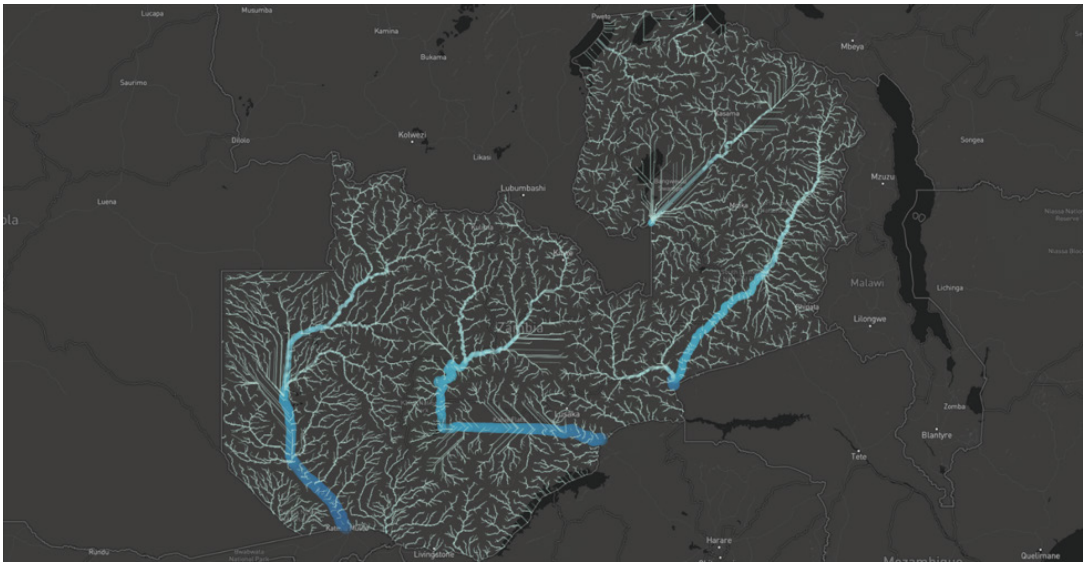


Figure 23 Channel Network of Zambia, vector rendered using Tableau; background map tiles courtesy © Mapbox, © OpenStreetMap

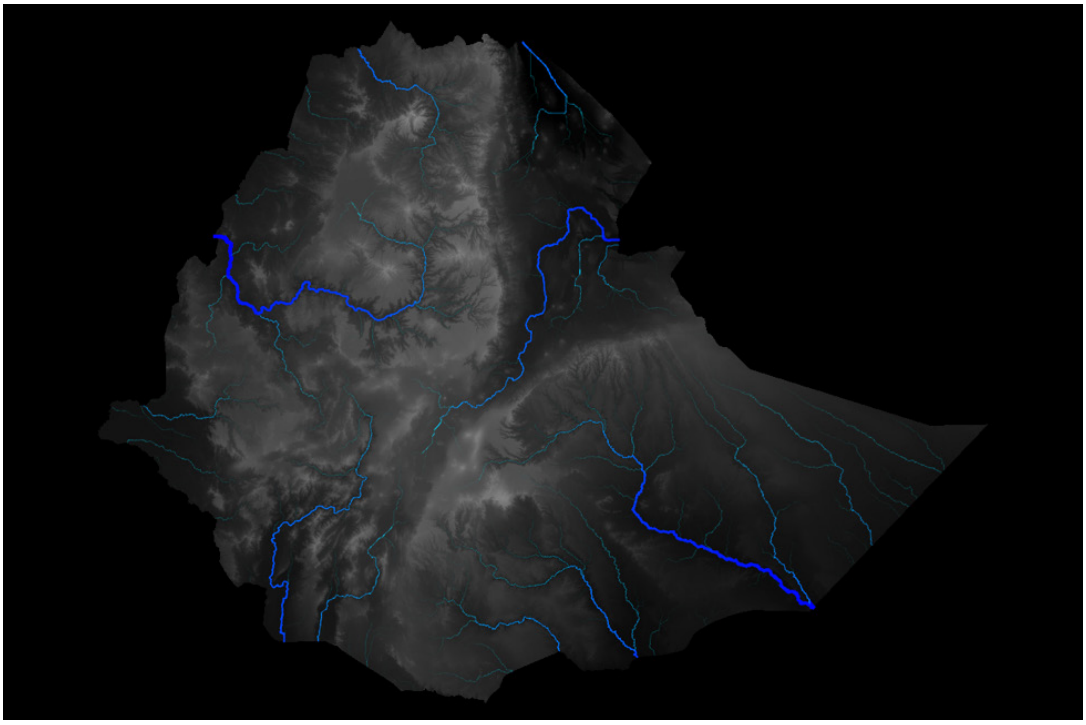


Figure 24 Channel Network of Ethiopia, polyline vector rendered using QGIS

Method

To obtain the Channel Network and Strahler Order, SAGA GIS (Python) Channel Network and Drainage Basins algorithm was executed, using the Filled DEM GeoTIFFs for Ethiopia, Senegal, and Zambia as the input elevation parameter respectively. This produces multiple outputs: a raster GeoTIFF of the Channel Network, a vector polyline of the channels, and a point shapefile of the junctions.

A vector array of the Strahler Order raster product was generated using the procedure mentioned hitherto.

Land Cover Classification

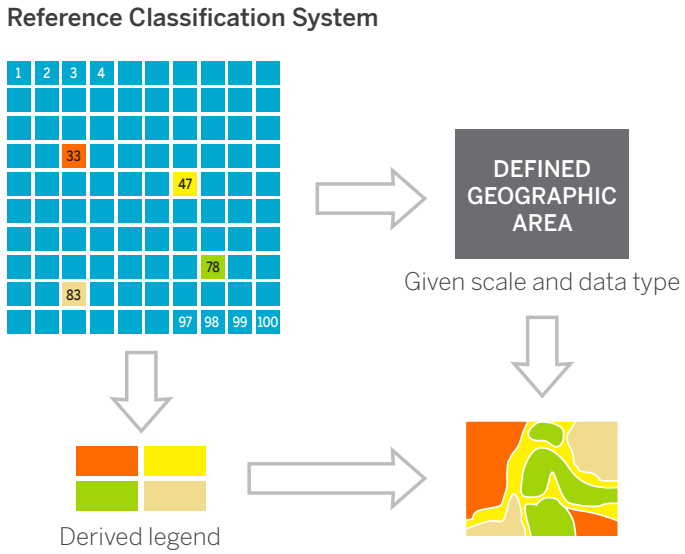


Figure 25 "Legend as application of a classification in a particular area" courtesy Natural Resources Management and Environment Department, Food and Agriculture Organization

Definition

Land cover is the observed (bio)physical cover on the earth's surface.

Classification is an abstract representation of the situation in the field using well-defined diagnostic criteria. (Gregorio, 2000)

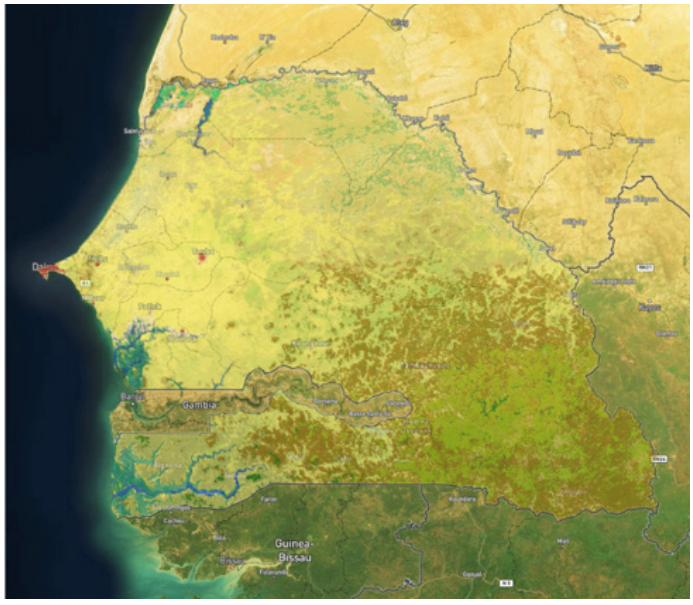


Figure 26 Land Cover of Senegal, vector rendered using Tableau; background map tiles courtesy © Mapbox, © Digital Globe, © OpenStreetMap

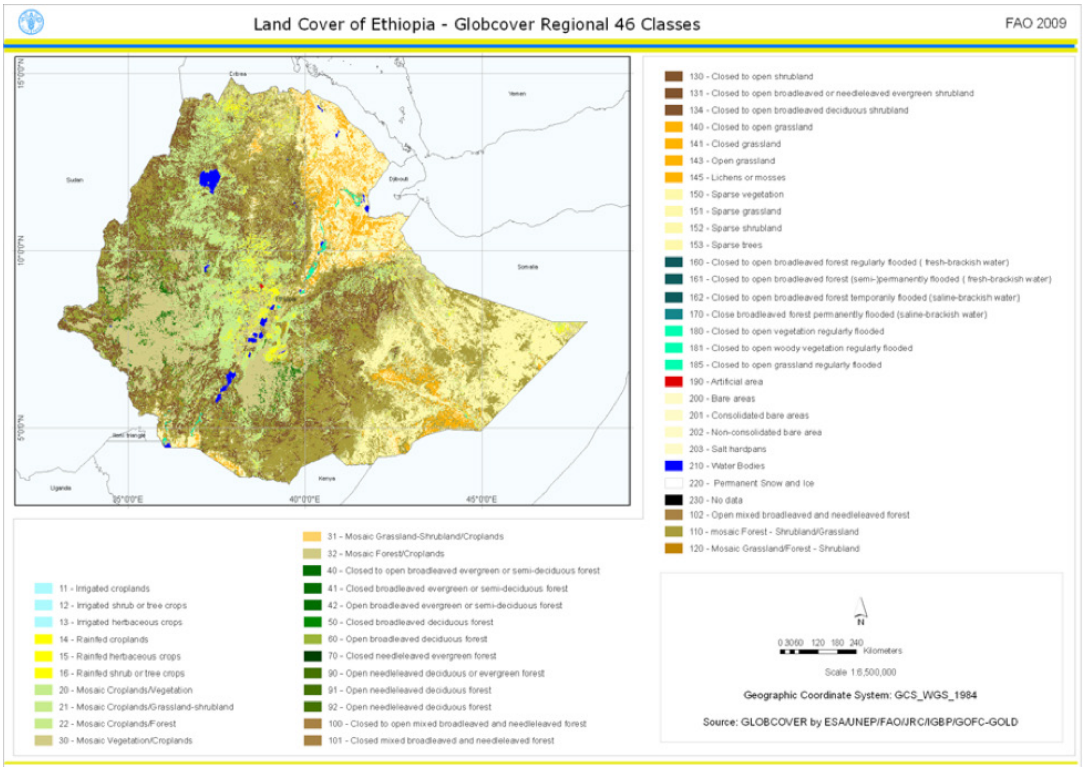


Figure 27 Land Cover of Senegal, courtesy FAO

Method

GlobCover (European Space Agency) vector polygon shapefiles from the FAO GLCN (Global Land Cover Network) repository were obtained. The 300m grids are enumerated by GRIDCODE that can be related to an accompanying data dictionary—a Microsoft Excel workbook—mapping GRIDCODE to labels. To map these labels, and to generate Boolean values, an Alteryx workflow was developed.

Each [GRIDCODE] was joined to the corresponding label on [VALUE] ("210" joined to "Water Bodies", "190" joined to "Artificial surfaces and associated areas (Urban Areas >50%), and then for each Grid code a conditional formula was developed; e.g. IF [GRIDCODE] = "210" THEN 1 ELSE 0 ENDIF.

The Microsoft Excel file also maps Red | Green | Blue values against each grid code, which was converted to a hex value, so each grid could then be encoded by color (for visualization purposes only— these values were not required in the model as input variables).

As before, the polygon centroids were resolved using the SAGA GIS (Python) command, and exported to CSVs for both Senegal, Ethiopia, and Zambia respectively.

Population Density

Definition

By integrating census, survey, satellite and administrative boundary datasets high resolution maps of population counts and densities have been provided by WorldPop. (Southampton, n.d.)

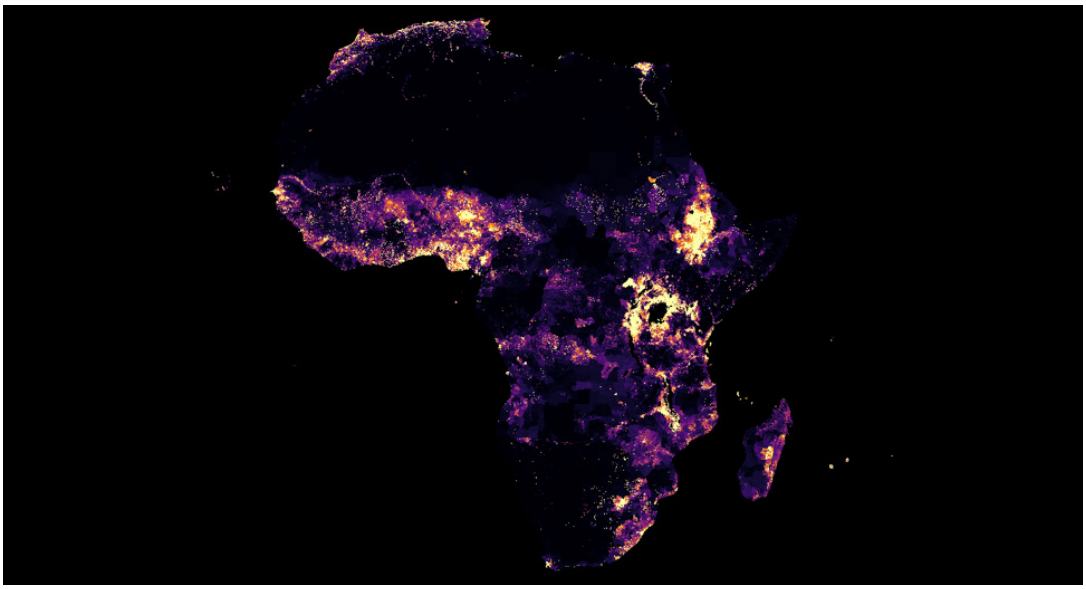


Figure 28 Population Density of Africa, 2015, raster GeoTIFF rendered in QGIS

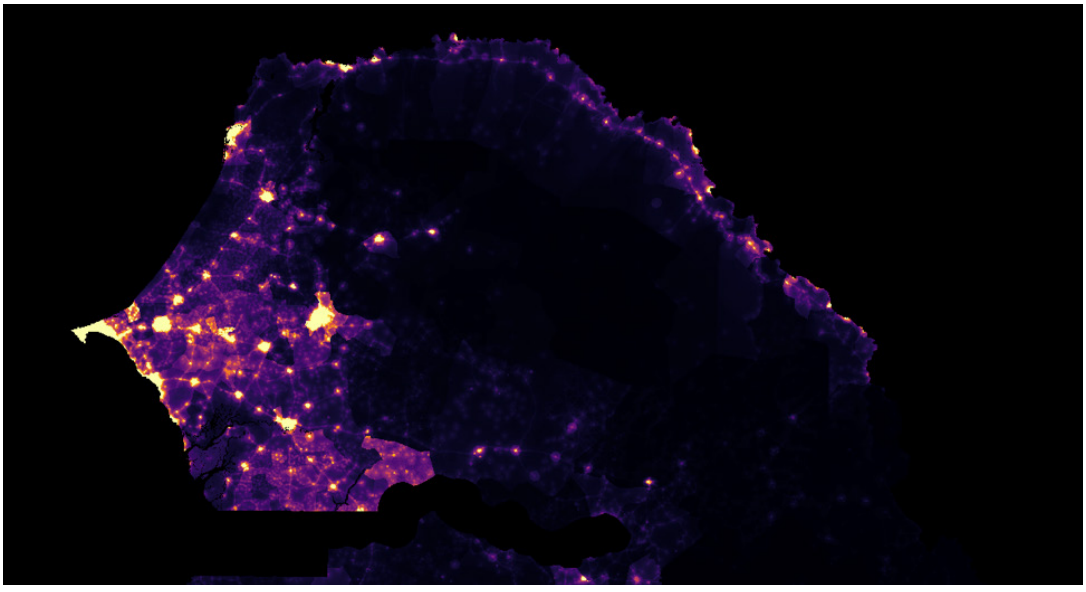


Figure 29 Population Density of Senegal, 2015, raster GeoTIFF rendered in QGIS

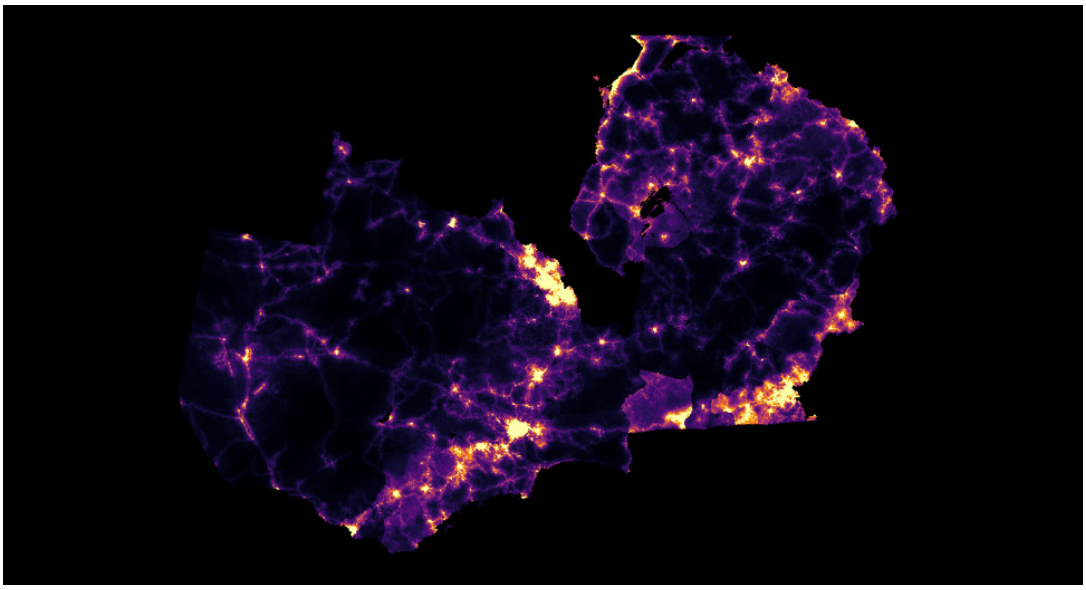


Figure 30 Population Density of Zambia, 2015, raster GeoTIFF rendered in QGIS

Method

GeoTIFF raster files were obtained from the WorldPop repository for years 2010, 2015, and 2020 for both Senegal and Zambia. The People Per Hectare (PPH) of the unadjusted United Nations estimates value files were chosen. For Ethiopia only years 2010 and 2015 were available.

All files were re-projected using the GDAL warp function in QGIS to WGS84 EPSG:4326. For each GeoTIFF, the GDAL command polygonize function was executed, resulting in six vector polygon shapefiles. To obtain an array of X (Longitude in decimal degrees), Y (Latitude in decimal degrees), and POP 2010 | POP 2015 | POP 2020, the three corresponding shapefiles were merged using MMQGIS, a plug-in tool for QGIS. The Ethiopia files were not merged.

The polygon centroids of the merged vector polygon shapefiles were resolved using SAGA GIS (Python) and exported to CSV.

To extrapolate the population density array, interpolated values for years in between 2010, 2015, and 2020 were resolved using an Alteryx workflow for Senegal and Zambia.

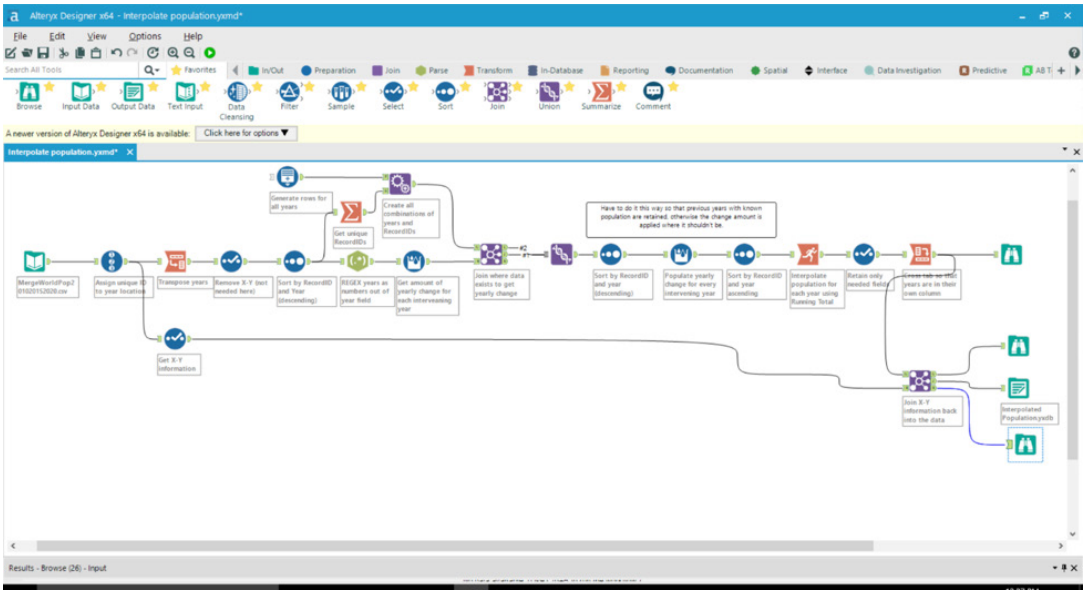
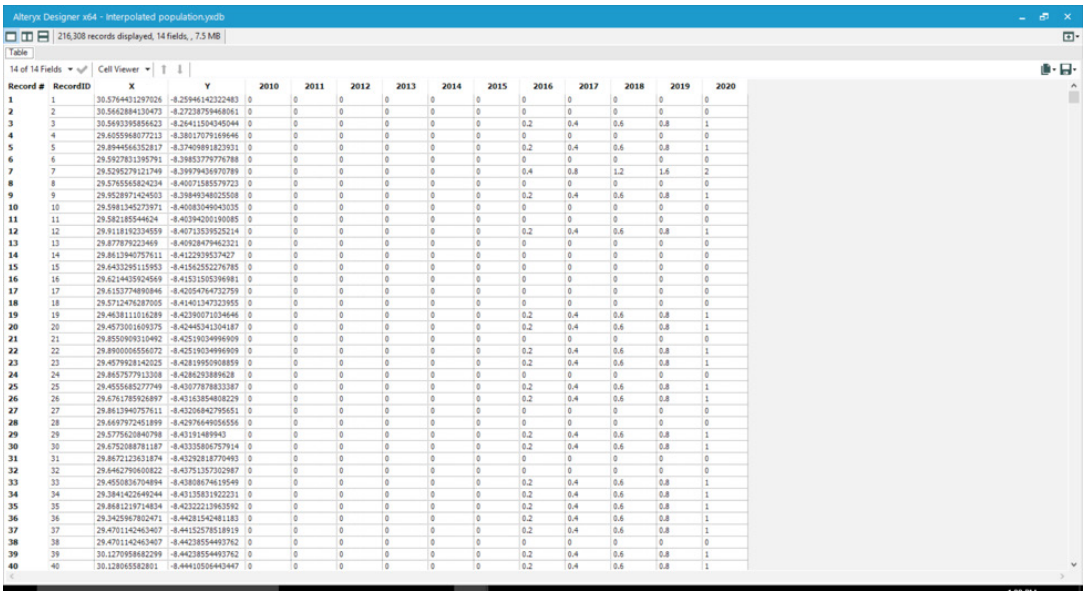


Figure 31 Alteryx Workflow: Interpolated Population Density, courtesy Philip Riggs



Record #	RecordID	X	Y	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
1	1	30.5764431297026	-8.25946142322483	0	0	0	0	0	0	0	0	0	0	0
2	2	30.5662884130473	-8.27238759468061	0	0	0	0	0	0	0	0	0	0	0
3	3	30.5693395856623	-8.26411504345044	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
4	4	29.6055968077213	-8.28012707919646	0	0	0	0	0	0	0	0	0	0	0
5	5	29.8944566352817	-8.37409891823931	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
6	6	29.5927831395791	-8.39853779776788	0	0	0	0	0	0	0	0	0	0	0
7	7	29.5295279121749	-8.39979436970789	0	0	0	0	0	0	0.4	0.8	1.2	1.6	2
8	8	29.5765565824234	-8.40071585579723	0	0	0	0	0	0	0	0	0	0	0
9	9	29.9528971424503	-8.39849348025508	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
10	10	29.5981345273971	-8.40083048040035	0	0	0	0	0	0	0	0	0	0	0
11	11	29.5821855444624	-8.40394200190085	0	0	0	0	0	0	0	0	0	0	0
12	12	29.9118192334559	-8.40713539525214	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
13	13	29.877879223469	-8.40928479462321	0	0	0	0	0	0	0	0	0	0	0
14	14	29.8613940757611	-8.41229395374 27	0	0	0	0	0	0	0	0	0	0	0
15	15	29.6433295115953	-8.41562552276785	0	0	0	0	0	0	0	0	0	0	0
16	16	29.6214435924569	-8.41531505396981	0	0	0	0	0	0	0	0	0	0	0
17	17	29.6153774890846	-8.42054764732759	0	0	0	0	0	0	0	0	0	0	0
18	18	29.5712476287005	-8.41401347323955	0	0	0	0	0	0	0	0	0	0	0
19	19	29.4638111016289	-8.42390071034646	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
20	20	29.4573001609375	-8.42445341304187	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
21	21	29.8550909310492	-8.42519034996909	0	0	0	0	0	0	0	0	0	0	0
22	22	29.8900006556072	-8.42519034996909	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
23	23	29.4579928142025	-8.42819950908859	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
24	24	29.8657577913308	-8.42862038896 28	0	0	0	0	0	0	0	0	0	0	0
25	25	29.4555685277749	-8.43077878833387	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
26	26	29.6761785926897	-8.43163854808229	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
27	27	29.8613940757611	-8.43206842795651	0	0	0	0	0	0	0	0	0	0	0
28	28	29.6697972451899	-8.42976649056556	0	0	0	0	0	0	0	0	0	0	0
29	29	29.5775620840798	-8.43191489943	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
30	30	29.6752088781187	-8.43335806757914	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
31	31	29.8672122631874	-8.43292818770493	0	0	0	0	0	0	0	0	0	0	0
32	32	29.6446279060082 2	-8.43751357302987	0	0	0	0	0	0	0	0	0	0	0
33	33	29.4550836704894	-8.43808674619549	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
34	34	29.3841422649244	-8.43135831922231	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
35	35	29.8681219714834	-8.42322213963592	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
36	36	29.3425967802471	-8.44281542481183	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
37	37	29.4701142463407	-8.44152578518919	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
38	38	29.4701142463407	-8.44238544493762	0	0	0	0	0	0	0	0	0	0	0
39	39	30.1270958682299	-8.44238554493762	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1
40	40	30.1280655828 01	-8.44410506443447	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1

Figure 32 Alteryx Output: Interpolated Population Density, courtesy Philip Riggs

Population Migration

Definition

Unconstrained gravity-type spatial interaction model (Alessandro Sorichetta, 2016), *otherwise known as an Origin-Destination (OD) Matrix, or node-link diagram.*

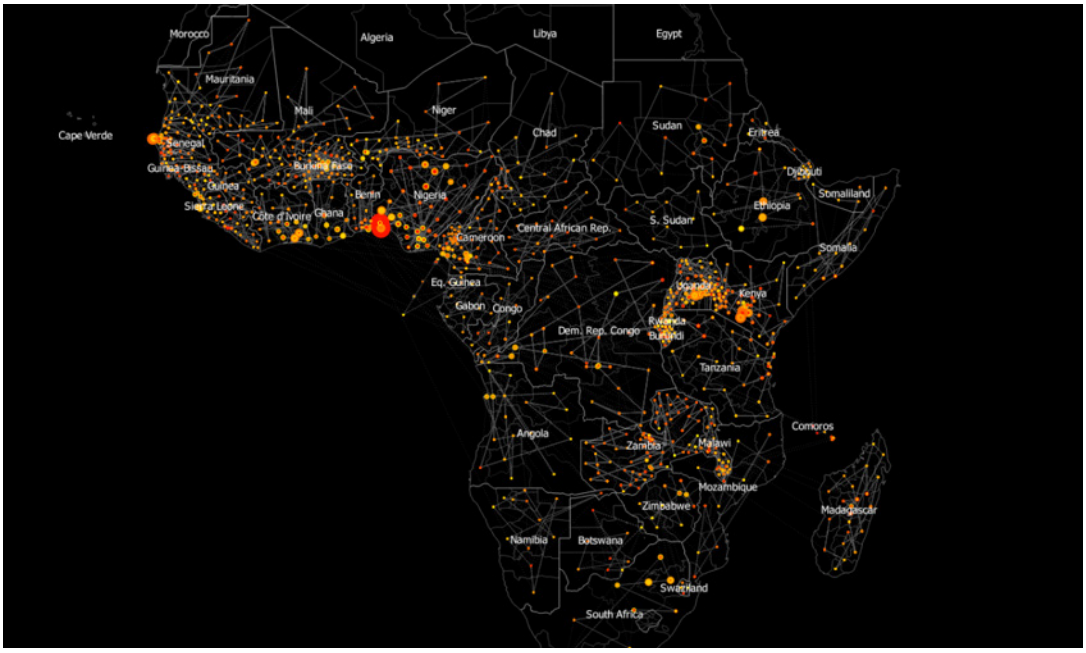
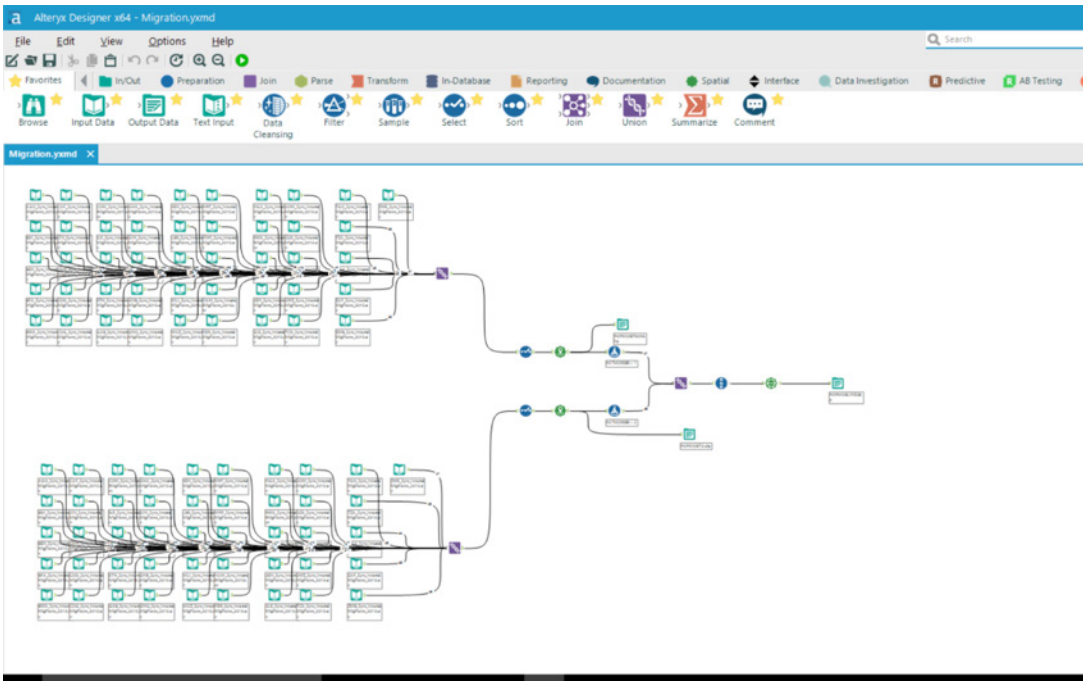


Figure 33 Estimated internal human migration flows between subnational administrative units for malaria endemic countries, 2010-2015, courtesy WorldPop, vector rendered in QGIS

Method

Multiple files in comma separated value format for each country were obtained from the WorldPop repository, and an Alteryx workflow was designed to union data within the files, ultimately producing two outputs: one that contains the latitude and longitude of the "From" nodes, and one that contains the latitude and longitude of the "To" nodes, with an integer value representing the number of people.

Figure 34 Alteryx Workflow: Migration



Multivariate Vector Array

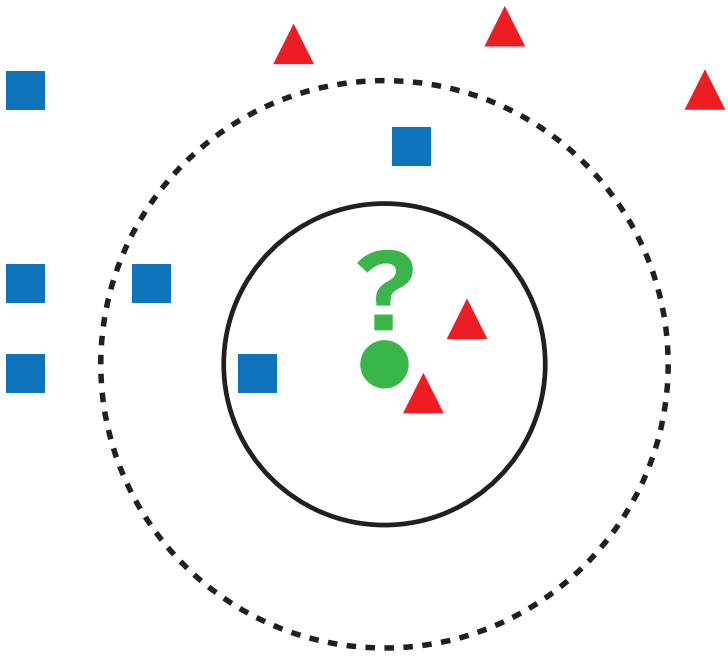


Figure 35 "Example of k-NN Classification" courtesy Wikipedia

Definition

Multiple variables are made available (using K-Nearest Neighbor) for simultaneous analysis. (Anderson, 1958)

Method

To extract the features of all the input variables into a single array, an Alteryx workflow was developed utilizing the Find Nearest tool which finds the result from the nearest Euclidean distance between two points. This method was chosen over the Spatial Match tool (also known as intersect) due to the size of the vector involved. This array is then stored in a cloud based in-memory database, EXASOL.

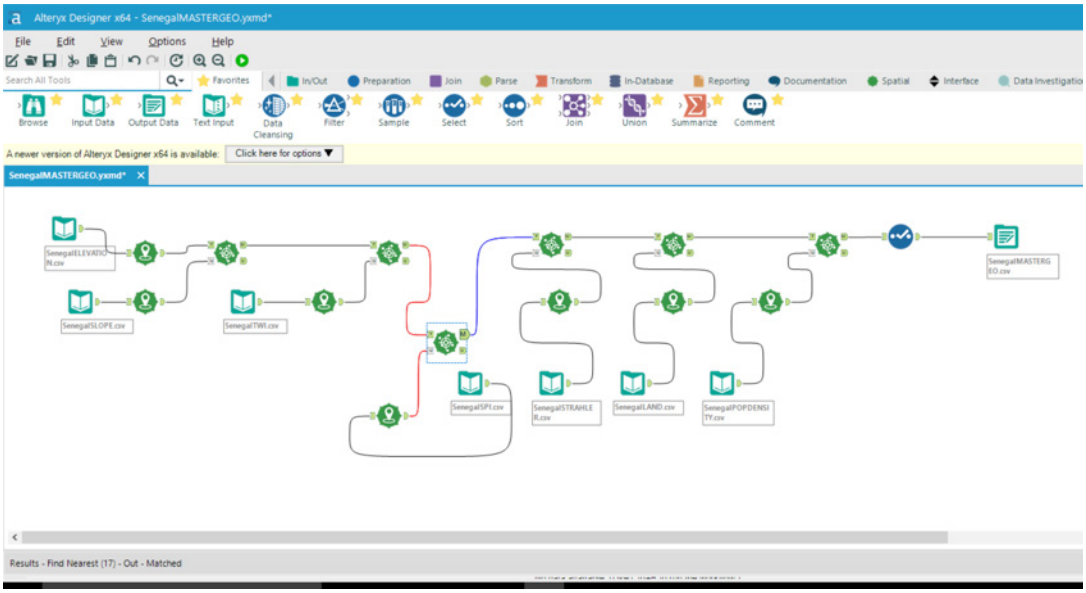
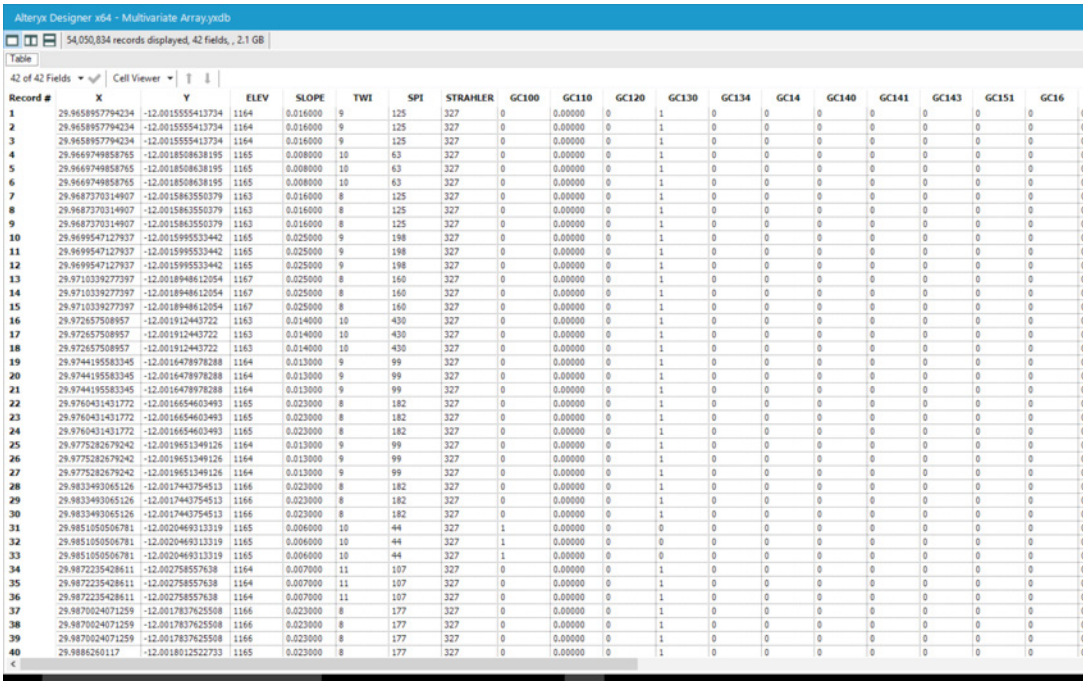


Figure 36 Alteryx Workflow: Find Nearest



Record #	X	Y	ELEV	SLOPE	TWI	SPI	STRAHLER	GC100	GC110	GC120	GC130	GC134	GC14	GC140	GC141	GC143	GC151	GC16
1	29.9658957794234	-12.0015555413734	1164	0.016000	9	125	327	0	0.00000	0	1	0	0	0	0	0	0	0
2	29.9658957794234	-12.0015555413734	1164	0.016000	9	125	327	0	0.00000	0	1	0	0	0	0	0	0	0
3	29.9658957794234	-12.0015555413734	1164	0.016000	9	125	327	0	0.00000	0	1	0	0	0	0	0	0	0
4	29.9669749858765	-12.0018508638195	1165	0.008000	10	63	327	0	0.00000	0	1	0	0	0	0	0	0	0
5	29.9669749858765	-12.0018508638195	1165	0.008000	10	63	327	0	0.00000	0	1	0	0	0	0	0	0	0
6	29.9669749858765	-12.0018508638195	1165	0.008000	10	63	327	0	0.00000	0	1	0	0	0	0	0	0	0
7	29.9687370314907	-12.0015863550379	1163	0.016000	8	125	327	0	0.00000	0	1	0	0	0	0	0	0	0
8	29.9687370314907	-12.0015863550379	1163	0.016000	8	125	327	0	0.00000	0	1	0	0	0	0	0	0	0
9	29.9687370314907	-12.0015863550379	1163	0.016000	8	125	327	0	0.00000	0	1	0	0	0	0	0	0	0
10	29.9699547127937	-12.0015995533442	1165	0.025000	9	198	327	0	0.00000	0	1	0	0	0	0	0	0	0
11	29.9699547127937	-12.0015995533442	1165	0.025000	9	198	327	0	0.00000	0	1	0	0	0	0	0	0	0
12	29.9699547127937	-12.0015995533442	1165	0.025000	9	198	327	0	0.00000	0	1	0	0	0	0	0	0	0
13	29.9710339277397	-12.0018948612054	1167	0.025000	8	160	327	0	0.00000	0	1	0	0	0	0	0	0	0
14	29.9710339277397	-12.0018948612054	1167	0.025000	8	160	327	0	0.00000	0	1	0	0	0	0	0	0	0
15	29.9710339277397	-12.0018948612054	1167	0.025000	8	160	327	0	0.00000	0	1	0	0	0	0	0	0	0
16	29.972657508957	-12.0019124443722	1163	0.014000	10	430	327	0	0.00000	0	1	0	0	0	0	0	0	0
17	29.972657508957	-12.0019124443722	1163	0.014000	10	430	327	0	0.00000	0	1	0	0	0	0	0	0	0
18	29.972657508957	-12.0019124443722	1163	0.014000	10	430	327	0	0.00000	0	1	0	0	0	0	0	0	0
19	29.9744195583345	-12.0016478978288	1164	0.013000	9	99	327	0	0.00000	0	1	0	0	0	0	0	0	0
20	29.9744195583345	-12.0016478978288	1164	0.013000	9	99	327	0	0.00000	0	1	0	0	0	0	0	0	0
21	29.9744195583345	-12.0016478978288	1164	0.013000	9	99	327	0	0.00000	0	1	0	0	0	0	0	0	0
22	29.9760431431772	-12.0016654603493	1165	0.023000	8	182	327	0	0.00000	0	1	0	0	0	0	0	0	0
23	29.9760431431772	-12.0016654603493	1165	0.023000	8	182	327	0	0.00000	0	1	0	0	0	0	0	0	0
24	29.9760431431772	-12.0016654603493	1165	0.023000	8	182	327	0	0.00000	0	1	0	0	0	0	0	0	0
25	29.9775282679242	-12.0019651349126	1164	0.013000	9	99	327	0	0.00000	0	1	0	0	0	0	0	0	0
26	29.9775282679242	-12.0019651349126	1164	0.013000	9	99	327	0	0.00000	0	1	0	0	0	0	0	0	0
27	29.9775282679242	-12.0019651349126	1164	0.013000	9	99	327	0	0.00000	0	1	0	0	0	0	0	0	0
28	29.9833493065126	-12.0017443754513	1166	0.023000	8	182	327	0	0.00000	0	1	0	0	0	0	0	0	0
29	29.9833493065126	-12.0017443754513	1166	0.023000	8	182	327	0	0.00000	0	1	0	0	0	0	0	0	0
30	29.9833493065126	-12.0017443754513	1166	0.023000	8	182	327	0	0.00000	0	1	0	0	0	0	0	0	0
31	29.9851050506781	-12.0020469313319	1165	0.006000	10	44	327	1	0.00000	0	0	0	0	0	0	0	0	0
32	29.9851050506781	-12.0020469313319	1165	0.006000	10	44	327	1	0.00000	0	0	0	0	0	0	0	0	0
33	29.9851050506781	-12.0020469313319	1165	0.006000	10	44	327	1	0.00000	0	0	0	0	0	0	0	0	0
34	29.9872235428611	-12.0027585557638	1164	0.007000	11	107	327	0	0.00000	0	1	0	0	0	0	0	0	0
35	29.9872235428611	-12.0027585557638	1164	0.007000	11	107	327	0	0.00000	0	1	0	0	0	0	0	0	0
36	29.9872235428611	-12.0027585557638	1164	0.007000	11	107	327	0	0.00000	0	1	0	0	0	0	0	0	0
37	29.9870024071259	-12.0017837625508	1166	0.023000	8	177	327	0	0.00000	0	1	0	0	0	0	0	0	0
38	29.9870024071259	-12.0017837625508	1166	0.023000	8	177	327	0	0.00000	0	1	0	0	0	0	0	0	0
39	29.9870024071259	-12.0017837625508	1166	0.023000	8	177	327	0	0.00000	0	1	0	0	0	0	0	0	0
40	29.9886260117	-12.0018012522733	1165	0.023000	8	177	327	0	0.00000	0	1	0	0	0	0	0	0	0

Figure 37 Alteryx Output: Multivariate Array

Voronoi of Facility Catchment Areas

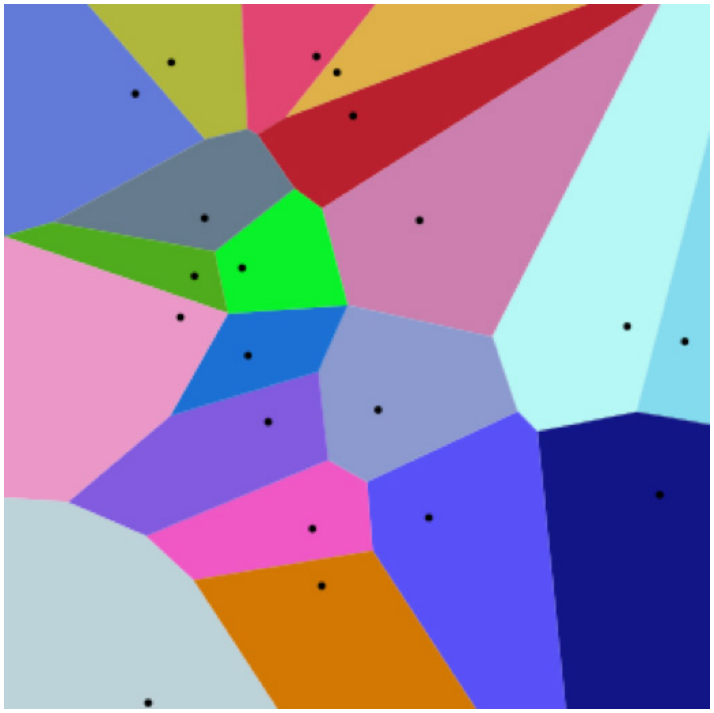


Figure 38 "20 Points and their Voronoi cells" courtesy Wikipedia

Definition

The partitioning of a plane with n points into convex polygons such that each polygon contains exactly one generating point, and every point in each polygon is closer to its generating point than to any other.

A particularly notable use of a Voronoi diagram was the analysis of the 1854 cholera epidemic in London, in which physician John Snow determined a strong correlation of deaths with proximity to an infected water pump on Broad Street. (Weisstein, n.d.)

Method

Point Location information of Health Facilities in Zambia, provided to PATH and MACEPA (The Malaria Control and Elimination Partnership in Africa) by the Zambian Ministry of Health and held in a DHIS 2 (District Health Information Software) database, were extracted and used as seeds to generate synthetic areas in a Voronoi. Each point is then intersected by a parent administrative boundary (District level for Zambia) with the product of the clipped intersection producing the output vector.

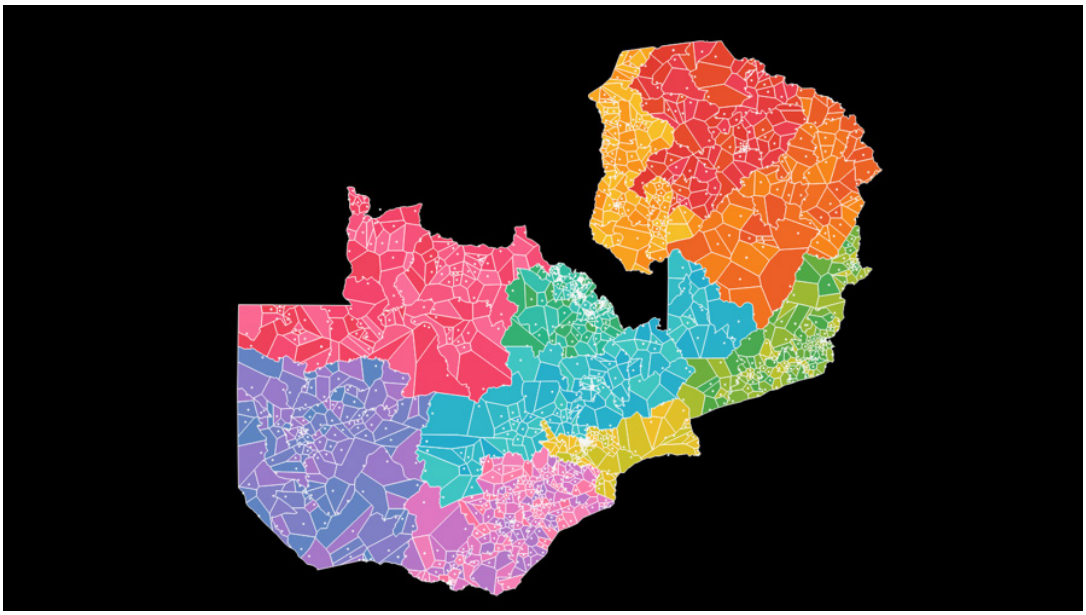


Figure 39 Alteryx Workflow: Facility Voronoi, courtesy Anya A'Hearn & Joe Mako

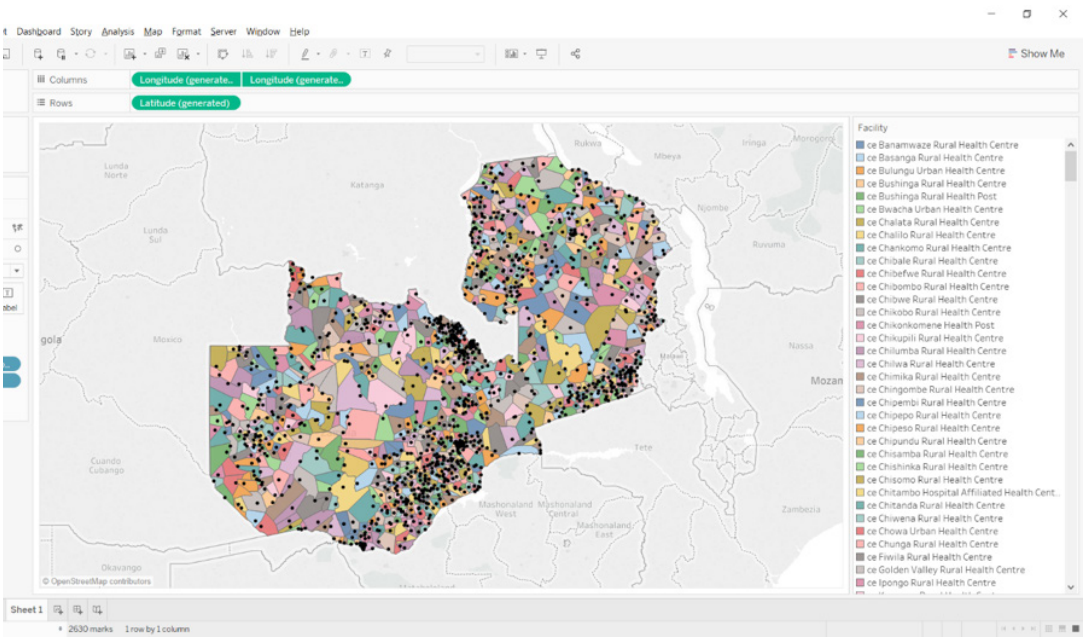


Figure 40 Alteryx Output: Facility Voronoi of Zambia, vector rendered using Tableau Software, courtesy Anya A'Hearn

Enriching Facilities with multivariate metrics

Definition

Each Health Facility has an administrative taxonomy above it (for example, District, Province, Country) and the latitude and longitude in decimal degrees recorded.

Method

To enrich each Facility with associated multivariate metrics, an Alteryx workflow was developed utilizing the Find Nearest tool as mentioned hitherto. The array containing the associated multivariate is queried from the cloud based in-memory database, EXASOL.

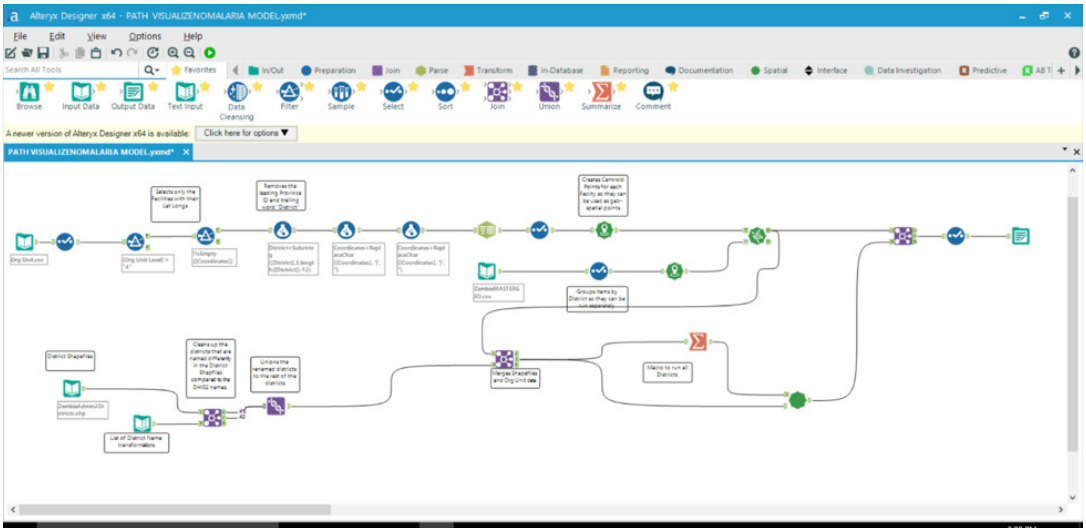


Figure 41 Alteryx Workflow: Joining Facility dimensions to multivariate metrics

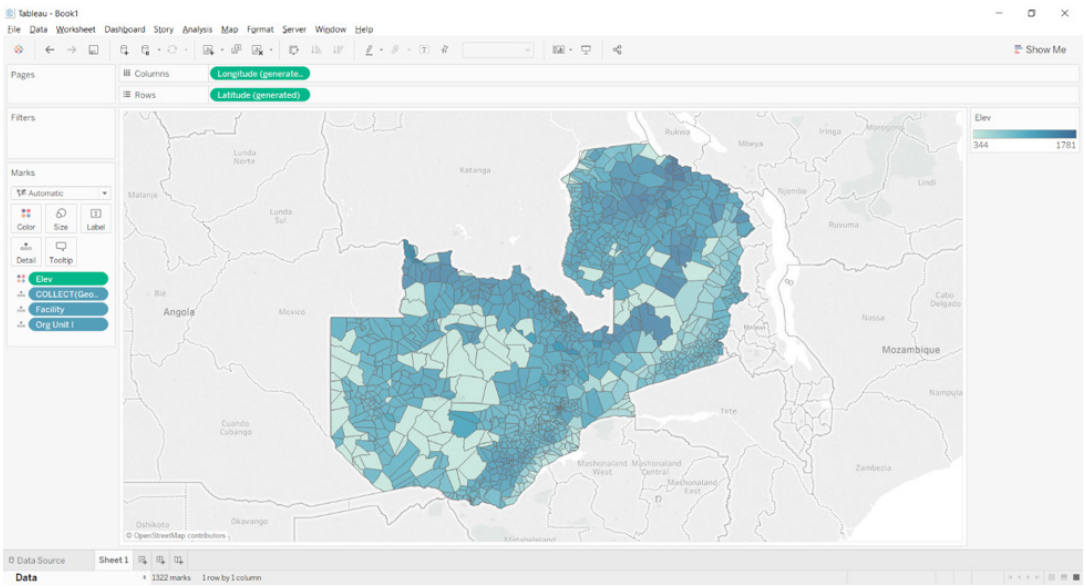


Figure 42 Alteryx Output: Facility Voronoi of Zambia, enriched with multivariate metrics (example Elevation), vector rendered using Tableau Software

Target Variable

Malaria Case

Definition

For Zambia, Malaria Cases are defined at the Facility level as the following:

- Total Confirmed Malaria (Positive Cases < 5 years + Positive Cases >= 5 years) yields the count [Malaria Cases]; and
- Step B: Total OPD Attendance (Total out-patient attendance) - Total Confirmed Malaria (Positive Cases < 5 years + Positive Cases >= 5 years) yields the count [not Malaria Cases].

Method

Two CSV files, exported from DHIS 2 Tables containing the temporal information (reporting cadence) and the Epidemiology records are (inner) joined by [Period ID] and loaded into the Alteryx workflow. Active Facilities report on a weekly cadence, so [Period Type] is filtered for "Weekly", and [Active] to equal "True".

Nota Bene: It is planned to automate the export from the DHIS 2 tables of these CSVs to a cloud based in-memory database, EXASOL, on a scheduled weekly cadence.

Each Facility is enriched with multivariate values, mentioned hitherto.

To generate a [Malaria Case] field, with a Boolean value of 1 for [Malaria Cases] and 0 of [not Malaria Cases], for each Facility weekly count, an Alteryx macro generates a row per count; exempli gratia:

Facility ID	Period Start Date	Malaria Cases	Non Malaria Cases
1234	1/1/2015	1	3

Resolves to:

Facility ID	Period Start Date	Malaria Case
1234	1/1/2015	1
1234	1/1/2015	0
1234	1/1/2015	0
1234	1/1/2015	0

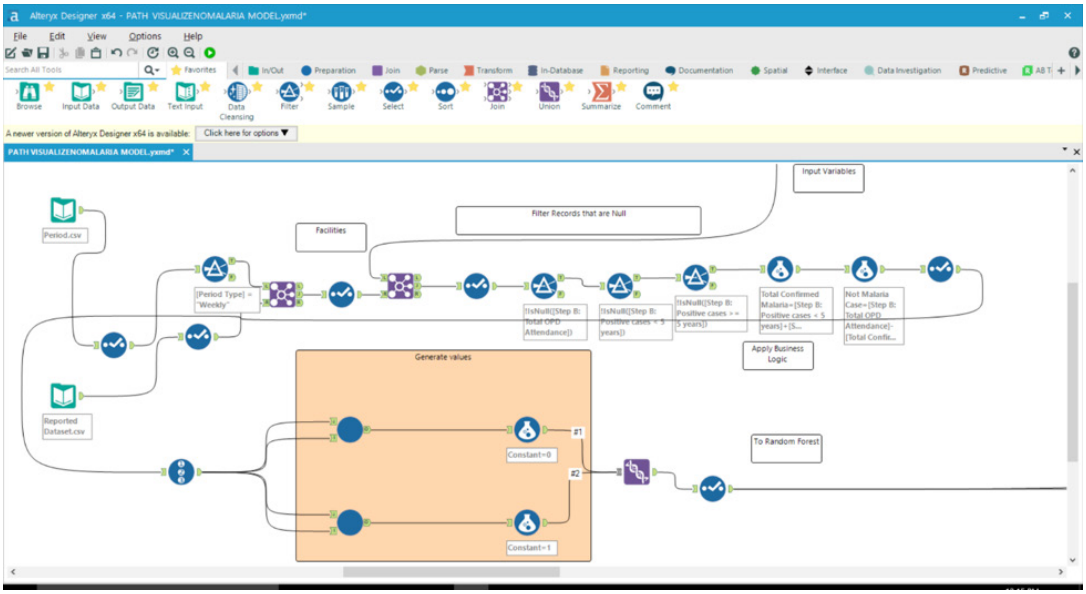
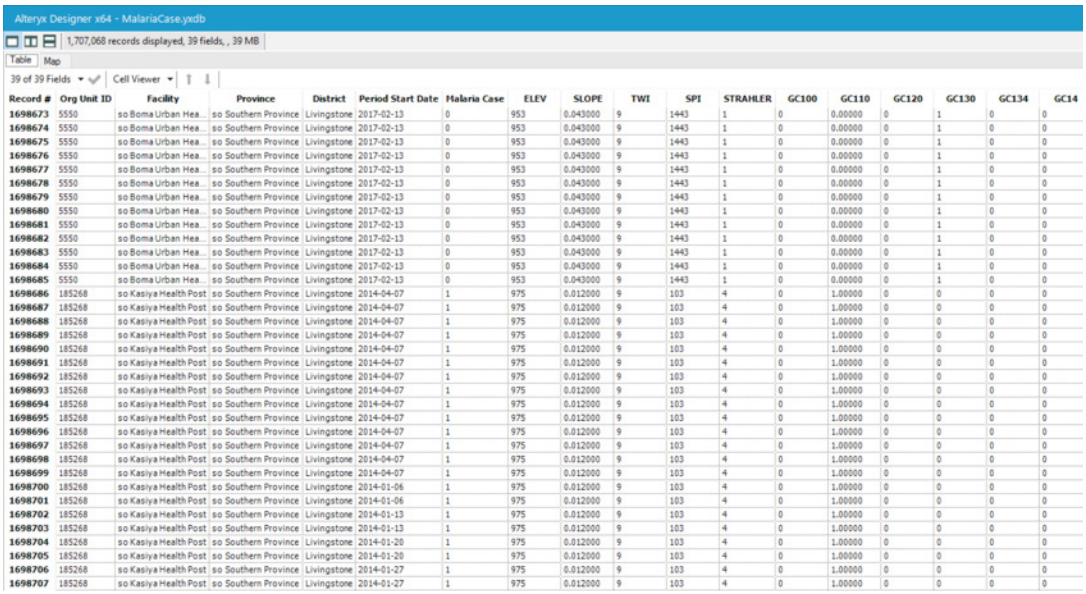


Figure 43 Alteryx Workflow: Target Variable Malaria Case Macro, courtesy Philip Riggs & Douglas Morris



Record #	Org Unit ID	Facility	Province	District	Period Start Date	Malaria Case	ELEV	SLOPE	TWI	SPI	STRAHLER	GC100	GC110	GC120	GC130	GC134	GC14
1698673	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698674	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698675	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698676	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698677	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698678	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698679	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698680	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698681	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698682	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698683	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698684	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698685	5550	so Boma Urban Hea...	so Southern Province	Livingstone	2017-02-13	0	953	0.043000	9	1443	1	0	0.00000	0	1	0	0
1698686	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698687	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698688	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698689	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698690	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698691	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698692	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698693	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698694	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698695	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698696	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698697	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698698	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698699	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-04-07	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698700	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-01-06	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698701	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-01-06	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698702	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-01-13	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698703	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-01-13	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698704	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-01-20	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698705	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-01-20	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698706	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-01-27	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0
1698707	185268	so Kasiya Health Post	so Southern Province	Livingstone	2014-01-27	1	975	0.012000	9	103	4	0	1.00000	0	0	0	0

Figure 44 Alteryx Output: Target Variable, Malaria Case

Random Forest

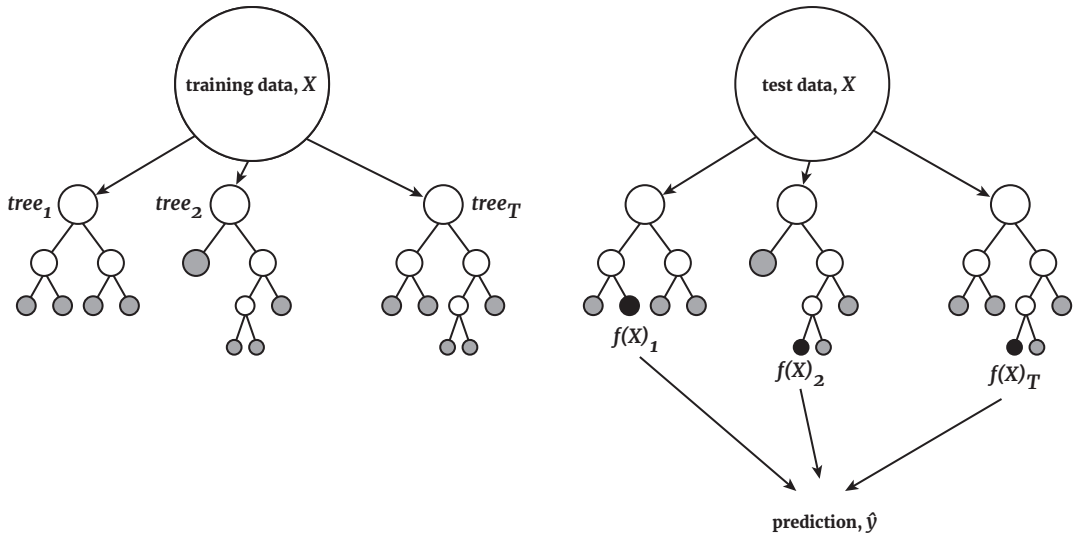


Figure 45 Conceptual Diagram of the Random Forest, courtesy The Journal of the Acoustical Society of America

Definition

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost (Freund and Schapire,1996)*, but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.* (Breiman, 2001)

Method

An Alteryx macro (Random Forest tool) was developed using a Graphical User Interface wrapper of the R package "randomForest" (Fortran original by Leo Breiman and Adele Cutler, 2015). Prior to invoking the tool, sampling occurs (80/20), then the input variables and target variables mentioned hitherto are entered as parameters. The fitted values are appended to the data stream by the Alteryx Score tool. The macro is batch executed per Health Facility.

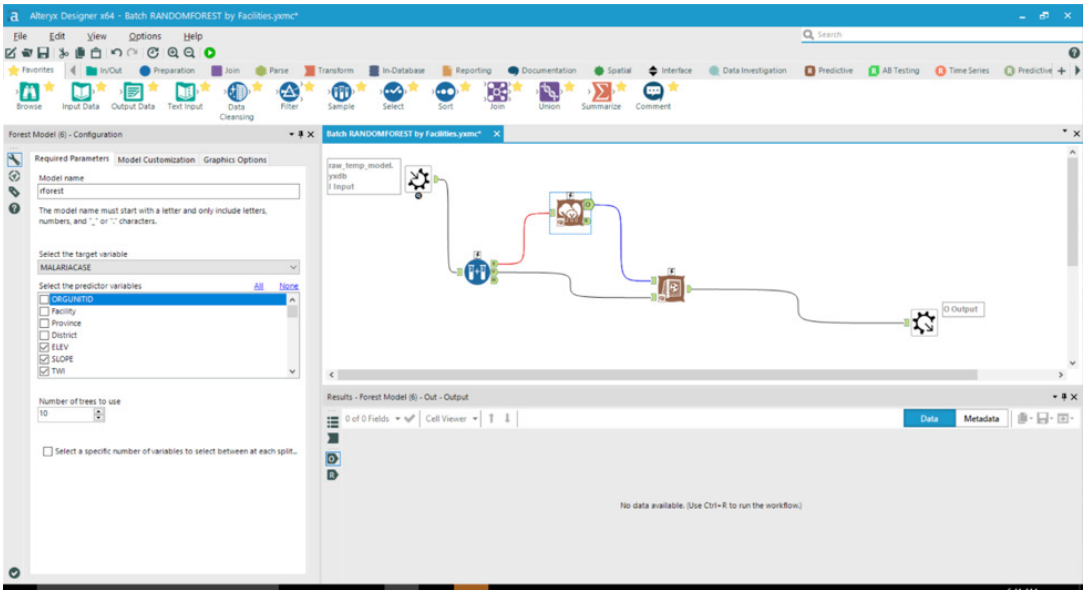
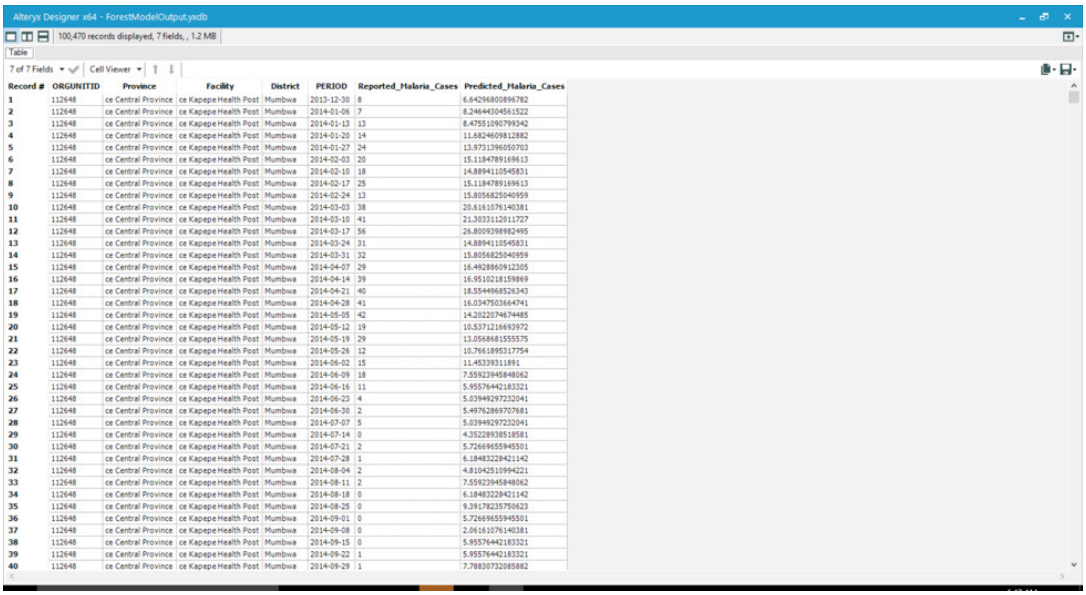


Figure 46 Alteryx Macro: Forest Model



Record #	ORGUNITID	Province	Facility	District	PERIOD	Reported_Malaria_Cases	Predicted_Malaria_Cases
1	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2013-12-30	8	6.64296800896782
2	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-01-06	7	8.24644304561522
3	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-01-13	13	8.47551090799342
4	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-01-20	14	11.6824609812882
5	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-01-27	24	13.9731396050703
6	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-02-03	20	15.1184789169613
7	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-02-10	18	14.8894110545831
8	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-02-17	25	15.1184789169613
9	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-02-24	13	15.8056825040959
10	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-03-03	38	20.6161076140381
11	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-03-10	41	21.3033112011727
12	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-03-17	56	26.8009398982495
13	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-03-24	31	14.8894110545831
14	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-03-31	32	15.8056825040959
15	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-04-07	29	16.4928860912305
16	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-04-14	39	16.9510218159869
17	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-04-21	40	18.5544968526343
18	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-04-28	41	16.0347503664741
19	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-05-05	42	14.2022074674485
20	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-05-12	19	10.5371216693972
21	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-05-19	29	13.0568681555575
22	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-05-26	12	10.7661895317754
23	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-06-02	15	11.4533931891
24	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-06-09	18	7.55923945848062
25	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-06-16	11	5.95576442183321
26	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-06-23	4	5.03949297232041
27	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-06-30	2	5.49762869707681
28	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-07-07	5	5.03949297232041
29	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-07-14	0	4.35228938518581
30	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-07-21	2	5.72669655945501
31	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-07-28	1	6.1848322842114
32	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-08-04	2	4.81042510994221
33	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-08-11	2	7.55923945848062
34	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-08-18	0	6.1848322842114
35	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-08-25	0	9.39178235750623
36	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-09-01	0	5.72669655945501
37	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-09-08	0	2.06161076140381
38	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-09-15	0	5.95576442183321
39	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-09-22	1	5.95576442183321
40	112648	ce Central Province	ce Kapepe Health Post	Mumbwa	2014-09-29	1	7.78830732085882

Figure 47 Alteryx Output: Forest Model

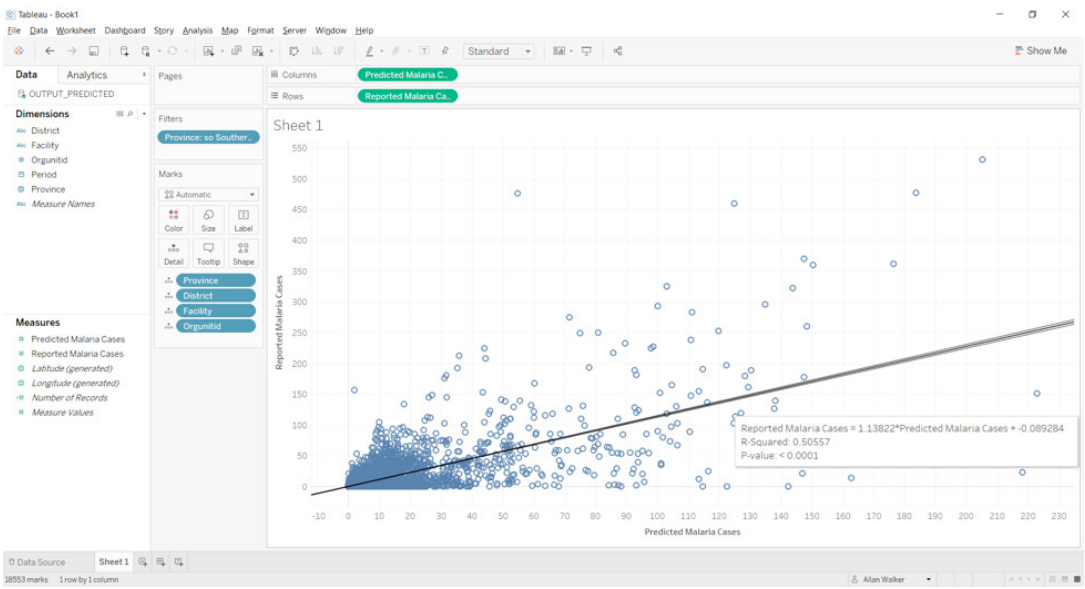


Figure 48 Initial Scatterplot of results of Random Forest for Southern Province, Zambia, rendered in Tableau Software

Forecast

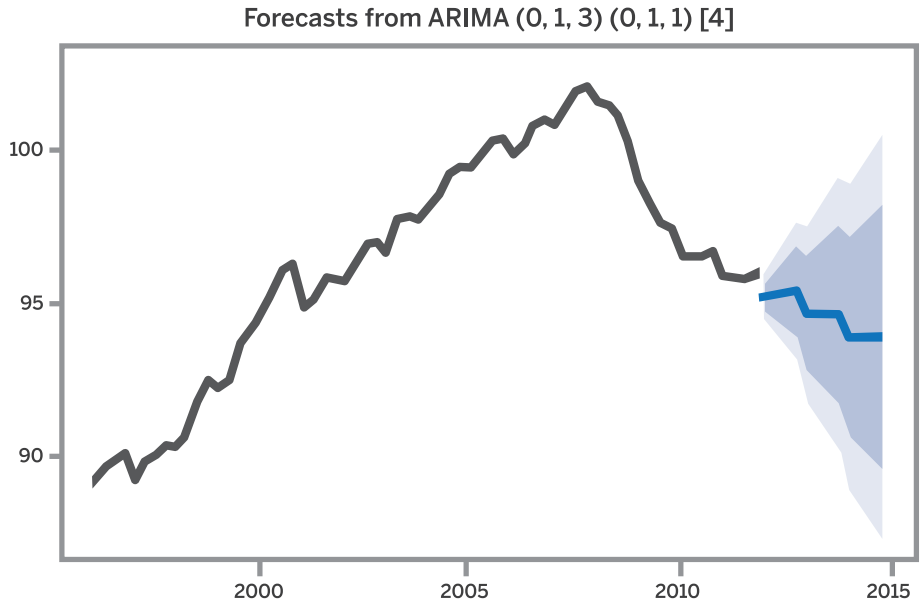


Figure 49 "Forecasts of the European retail trade index data using an ARIMA model" reference Forecasting: principles and practice, courtesy Rob J Hyndman & George Athanasopoulos

Definition

ARIMA models are, in theory, the most general class of models for forecasting a time series which can be made to be "stationary" by differencing (if necessary), perhaps in conjunction with nonlinear transformations such as logging or deflating (if necessary). A random variable that is a time series is stationary if its statistical properties are all constant over time. A stationary series has no trend, its variations around its mean have a constant amplitude, and it wiggles in a consistent fashion, i.e., its short-term random time patterns always look the same in a statistical sense. The latter condition means that its autocorrelations (correlations with its own prior deviations from the mean) remain constant over time, or equivalently, that its power spectrum remains constant over time. A random variable of this form can be viewed (as usual) as a combination of signal and noise, and the signal (if one is apparent) could be a pattern of fast or slow mean reversion, or sinusoidal oscillation, or rapid alternation in sign, and it could also have a seasonal component. An ARIMA model can be viewed as a "filter" that tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts. (Nau, 2017)

Method

An Alteryx macro was developed using a Graphical User Interface wrapper of the R package "forecast" (Rob Hyndman [aut, 2017). The ARIMA tool has the target variable of [Reported Malaria Cases] and co-variable of [Predicted Malaria Cases]. The Alteryx tool "TS Covariate Forecast" then appends the data stream with the forecast values and confidence bounds.

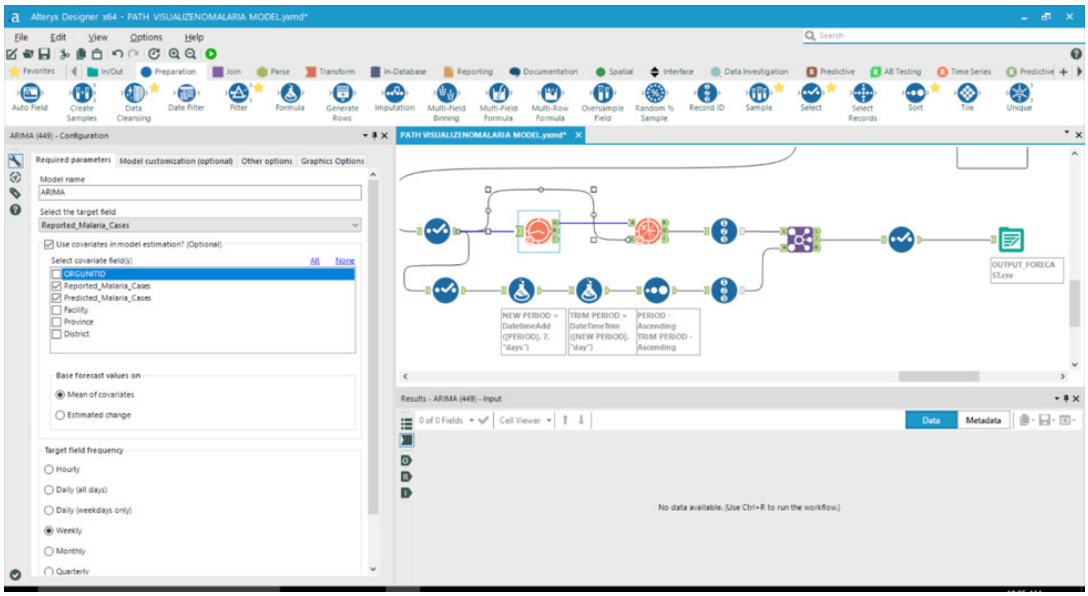
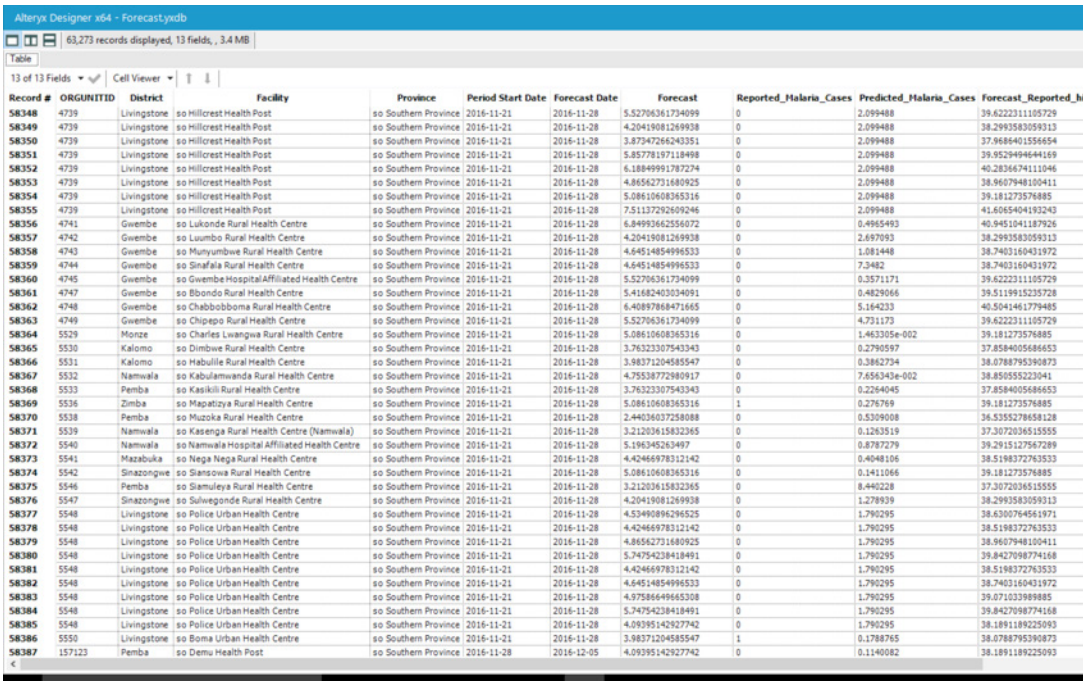


Figure 50 Alteryx Workflow "ARIMA and TS Covariate Forecast"



Alteryx Designer x64 - Forecast.yxdb

63,273 records displayed, 13 fields, , 3.4 MB

Table

13 of 13 Fields | Cell Viewer

Record #	ORGUNITID	District	Facility	Province	Period Start Date	Forecast Date	Forecast	Reported_Malaria_Cases	Predicted_Malaria_Cases	Forecast_Reported_h
58348	4739	Livingstone	so Hillcrest Health Post	so Southern Province	2016-11-21	2016-11-28	5.52706361734099	0	2.099488	39.6222311105729
58349	4739	Livingstone	so Hillcrest Health Post	so Southern Province	2016-11-21	2016-11-28	4.20419081269938	0	2.099488	38.2993583059313
58350	4739	Livingstone	so Hillcrest Health Post	so Southern Province	2016-11-21	2016-11-28	3.87347266243351	0	2.099488	37.9686401556654
58351	4739	Livingstone	so Hillcrest Health Post	so Southern Province	2016-11-21	2016-11-28	5.85778197118498	0	2.099488	39.9529494644169
58352	4739	Livingstone	so Hillcrest Health Post	so Southern Province	2016-11-21	2016-11-28	6.18849991787274	0	2.099488	40.2836674111046
58353	4739	Livingstone	so Hillcrest Health Post	so Southern Province	2016-11-21	2016-11-28	4.86562731680925	0	2.099488	38.9607948100411
58354	4739	Livingstone	so Hillcrest Health Post	so Southern Province	2016-11-21	2016-11-28	5.08610608365316	0	2.099488	39.181273576885
58355	4739	Livingstone	so Hillcrest Health Post	so Southern Province	2016-11-21	2016-11-28	7.51137292609246	0	2.099488	41.6065404193243
58356	4741	Gwembe	so Lukonde Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	6.84993662556072	0	0.4965493	40.9451041187926
58357	4742	Gwembe	so Luumbo Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	4.20419081269938	0	2.697093	38.2993583059313
58358	4743	Gwembe	so Munyumbwe Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	4.64514854996533	0	1.081448	38.7403160431972
58359	4744	Gwembe	so Sinafala Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	4.64514854996533	0	7.3482	38.7403160431972
58360	4745	Gwembe	so Gwembe Hospital Affiliated Health Centre	so Southern Province	2016-11-21	2016-11-28	5.52706361734099	0	0.3571171	39.6222311105729
58361	4747	Gwembe	so Bbondo Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	5.41682403034091	0	0.4829066	39.5119915235728
58362	4748	Gwembe	so Chabbobboma Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	6.40897868471665	0	5.164233	40.5041461779485
58363	4749	Gwembe	so Chipepo Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	5.52706361734099	0	4.731173	39.6222311105729
58364	5529	Monze	so Charles Lwangwa Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	5.08610608365316	0	1.463305e-002	39.181273576885
58365	5530	Kalomo	so Dimbwe Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	3.76323307543343	0	0.2790597	37.8584005686653
58366	5531	Kalomo	so Habulile Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	3.98371204585547	0	0.3862734	38.0788795390873
58367	5532	Namwala	so Kabulamwanda Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	4.75538772980917	0	7.656343e-002	38.850555223041
58368	5533	Pemba	so Kasikili Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	3.76323307543343	0	0.2264045	37.8584005686653
58369	5536	Zimba	so Mapatizya Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	5.08610608365316	1	0.276769	39.181273576885
58370	5538	Pemba	so Muzoka Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	2.44036037258088	0	0.5309008	36.5355278658128
58371	5539	Namwala	so Kasenga Rural Health Centre (Namwala)	so Southern Province	2016-11-21	2016-11-28	3.21203615832365	0	0.1263519	37.3072036515555
58372	5540	Namwala	so Namwala Hospital Affiliated Health Centre	so Southern Province	2016-11-21	2016-11-28	5.196345263497	0	0.8787279	39.2915127567289
58373	5541	Mazabuka	so Nega Nega Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	4.42466978312142	0	0.4048106	38.5198372763533
58374	5542	Sinazongwe	so Siansowa Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	5.08610608365316	0	0.1411066	39.181273576885
58375	5546	Pemba	so Siamuleya Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	3.21203615832365	0	8.440228	37.3072036515555
58376	5547	Sinazongwe	so Sulwegonde Rural Health Centre	so Southern Province	2016-11-21	2016-11-28	4.20419081269938	0	1.278939	38.2993583059313
58377	5548	Livingstone	so Police Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	4.53490896296525	0	1.790295	38.6300764561971
58378	5548	Livingstone	so Police Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	4.42466978312142	0	1.790295	38.5198372763533
58379	5548	Livingstone	so Police Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	4.86562731680925	0	1.790295	38.9607948100411
58380	5548	Livingstone	so Police Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	5.74754238418491	0	1.790295	39.8427098774168
58381	5548	Livingstone	so Police Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	4.42466978312142	0	1.790295	38.5198372763533
58382	5548	Livingstone	so Police Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	4.64514854996533	0	1.790295	38.7403160431972
58383	5548	Livingstone	so Police Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	4.97586649665308	0	1.790295	39.071033989885
58384	5548	Livingstone	so Police Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	5.74754238418491	0	1.790295	39.8427098774168
58385	5548	Livingstone	so Police Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	4.09395142927742	0	1.790295	38.1891189225093
58386	5550	Livingstone	so Boma Urban Health Centre	so Southern Province	2016-11-21	2016-11-28	3.98371204585547	1	0.1788765	38.0788795390873
58387	157123	Pemba	so Demu Health Post	so Southern Province	2016-11-28	2016-12-05	4.09395142927742	0	0.1140082	38.1891189225093

Figure 51 Alteryx Output: Forecast of Malaria Cases using Recorded and Predicted Values

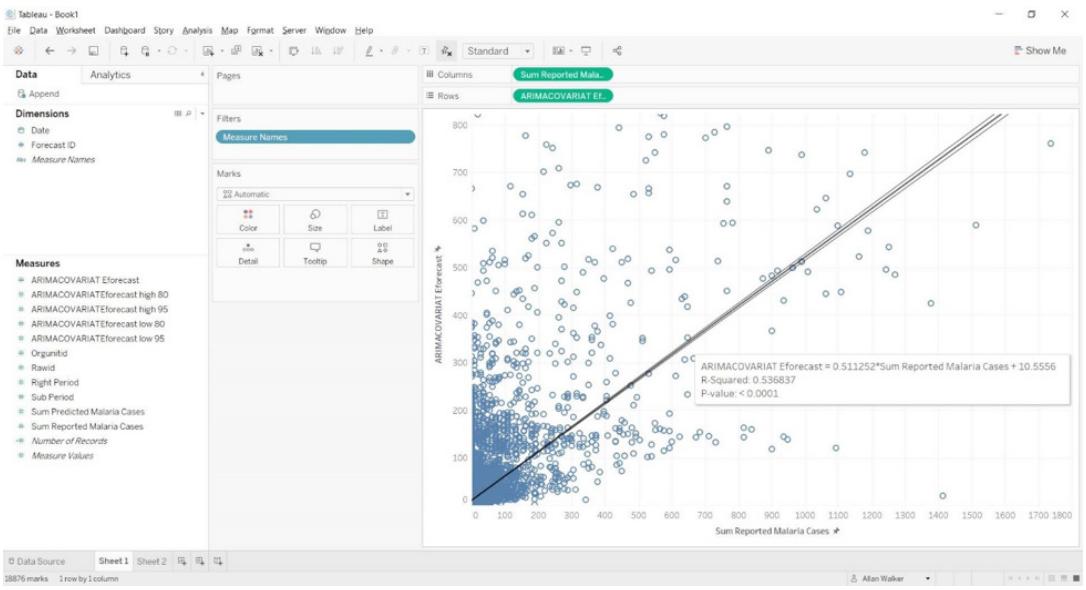


Figure 52 Initial Scatterplot of results of ARIMA for Southern Province, Zambia, rendered in Tableau Software

Conclusions

While initial Results of the workflow are promising, with good model fitting of prediction to recorded malaria cases, the author recognizes that the introduction of further input variables such as population mobility, meteorological, spray, net, and drug administration data could significantly improve the random forest model. The random model could also gain improvements by further configuration by a subject domain expert. The same applies to the ARIMA forecast model.

The cadence of the workflow schedule will also be important for meaningful results, as there is a lag between [Period Start Date] and the records being updated and finalized.

Archiving output and comparing could also provide valuable insight into the fidelity of the model, and allow for tuning/re-configuration.

Works Cited

Alessandro Sorichetta, T.J.-S. (2016). Africa > Whole Continent > Internal Migration Flows 2010. Tetrieved from WoldPop: http://web.archive.org/web/20040321033433/http://www.uwsp.edu/geo/faculty/ritter/glossary/a_d/drainage_basin.html

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley.

Breiman, L. (2001). RANDOM FORESTS. Retrieved from www.stat.berkeley.edu: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Fortran original by Leo Breiman and Adele Cutler, R. p. (2015). Package "randomForest". Retrieved from Breiman and Cutler's Random Forests for Classification and Regression: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Gallant, J. P. (2000). Digital Terrain Analysis. Retrieved from http://johnwilson.usc.edu: http://johnwilson.usc.edu/wp-content/uploads/2016/05/2000-Wilson-Gallant-Terrain-Anaylsis-Chapter-1.pdf

Gregorio, A. D. (2000). Land Cover Classification System (LCCS). Retrieved from http://www.fao.org: http://www.fao.org/docrep/003/x0596e/X0596e00.htm

Guard, U. S. (n.d.). Situation Awareness. Retrieved from United States Coast Guard: https://www.uscg.mil/auxiliary/training/tct/chap5.pdf

J., H. G., & F., G. M. (1997). "Modeling the uncertainty of slope and aspect estimates derived from spatial databases". Geographical Analysis. http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1997.tb00944.x/abstract

Justin M Cohen, K. C. (2010). Local topographic wetness indices predict household malaria risk better than land-use and land-cover in the western Kenya highlands. Retrieved from PMC2993734: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2993734/

Nau, R. (2017). ARIMA models for time series forecasting. Retrieved from Statistical forecasting: notes on regression and time series analysis: http://people.duke.edu/~rnau/411arim.htm

Nmor JC, S. T. (2013). Topographic models for predicting malaria vector breeding habitats: potential tools for vector control managers. Retrieved from Parasit Vectors: https://www.ncbi.nlm.nih.gov/pubmed/23324389

R. Sørensen, U. Z. (2006). On the calculation of the topographic wetness index: evaluation of different methods based on field observations. Retrieved from http://www.hydrol-earth-syst-sci.net: http://www.hydrol-earth-syst-sci.net/10/101/2006/hess-10-101-2006.pdf

Ryndman [au, c. c.-W. (2017). Package 'forecast'. Retrieved from Forecasting Functions for Time Series and Linear Models: https://cran.r-project.org/web/packages/forecast/forecast.pdf

Southampton, G. I. (n.d.). WorldPop. Retrieved from WorldPop: http://www.worldpop.org.uk

SRTM 90m Digital Elevation Database v4.1. (n.d.). Retrieved from http://www.cgiar-csi.org: http://www.cgiar-csi.org/data/srtm-90m-digital-elevation-database-v4-1

STRAHLER, A. N. (1952). HYPSOMETRIC (AREA-ALTITUDE) ANALYSIS OF EROSIONAL TOPOGRAPHY. Retrieved from Geological Society of America Bulletin: http://www.unc.edu/courses/2010spring/geog/591/001/students/nmey13/GEOL483/Lab5/pdfs/Strahler_1952_hypsometry.pdf

Weisstein, E. (n.d.). Voronoi Diagram. Retrieved from MathWorld--A Wolfram Web Resourc: http://mathworld.wolfram.com/VoronoiDiagram.html

Wisconsin–Stevens, U. o. (n.d.). The Physical Environment. "drainage basin". Retrieved from http://web.archive.org: http://web.archive.org/web/20040321033433/http://www.uwsp.edu/geo/faculty/ritter/glossary/a_d/drainage_basin.html

About Tableau

Tableau helps people transform data into actionable insights that make an impact. Easily connect to data stored anywhere, in any format. Quickly perform ad hoc analyses that reveal hidden opportunities. Drag and drop to create interactive dashboards with advanced visual analytics. Then share across your organization and empower teammates to explore their perspective on data. From global enterprises to early-stage startups and small businesses, people everywhere use Tableau's analytics platform to see and understand their data.