MIDLANDS STATE UNIVERSITY

FACULTY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF SURVEYING AND GEOMATICS

ASSESSING THE ACCURACY OF GLOBAL POPULATION DISTRIBUTION DATASETS FOR MATERNAL AND PERINATAL HEALTH APPLICATIONS

A dissertation submitted to the Faculty of Science and Technology, Department of Surveying and Geomatics at the Midlands State University in partial fulfilment of the requirements for the award of a Bachelor of Science Honours Degree in Surveying and Geomatics.

By

YOLISA PRUDENCE DUBE

(R122426H)

SUPERVISOR: Dr P.T. MAKANGA

DECLARATION

I, **Yolisa Prudence Dube** hereby declare that I am the sole author of this dissertation. I comprehend the nature of plagiarism, and I am cognisant of the university’s policies on this matter.

Signature ……………………………………………...

Date…………………………………………………….

APPROVAL FORM

The undersigned people certify that they read and recommend Midlands State University to accept a dissertation entitled, "Assessing the accuracy of global population distribution datasets for maternal and perinatal health applications" by Yolisa Prudence Dube in partial fulfilment of Bachelor of Science Honours Degree in Surveying and Geomatics.

Supervisor: ……………………. Signature………………… Date……. /………/…… 2016

Student: ………………………. Signature………………… Date……. /………/……2016

Chairperson: …………………... Signature……………....... Date……/………/……. 2016

RELEASE FORM

Name of student:	Yolisa Prudence Dube
Dissertation title:	Assessing the accuracy of global population distribution datasets for maternal and perinatal health applications
Degree title:	BSc Honours in Surveying and Geomatics
Year:	2016
Contact address:	1 Catalina Bay Cnr Galway & Joubert Street Germiston South Gauteng South Africa

Permission is hereby granted to the Midlands State University Library to produce single copies of this dissertation and to lend such copies for private, scholarly or scientific research purpose only.

The author reserves other publication rights. Neither the dissertation nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

SIGNED: ……………………………………

DATE: ……………………………………….

DEDICATION

I dedicate this dissertation to my lovely mother Ms Nonkumbulo Adonis-Maduna and brothers Clayton and Clarence Dube, who have been my pillar of strength and the reason I have made it this far. A special dedication to the memory of my late father Mr E. Dube who laid down the foundation for me and made me love and appreciate education.

ACKNOWLEDGEMENTS

I would like to thank the Lord for guiding me throughout my studies. My greatest gratitude goes to my supervisor Dr. P. T. Makanga for his encouragement and guidance throughout this dissertation. My gratitude is also extended to my mother Ms N. Adonis-Maduna for her support and contribution the success of this dissertation. I would also like to thank the staff of the Department of Surveying and Geomatics for their assistance in making this dissertation a success.

ABSTRACT

Global population distribution datasets have been used in a lot of studies and research programmes including public health research due to their availability and large scale geographical coverage. Their increasing application in maternal and perinatal studies has increased the implications of the data extracted from the datasets. The newly developed SDGs with very high expectations in terms of deliverables in the health care sector require high quality data which reveals the heterogeneity existing at subnational levels. These datasets as sources of data therefore need to be cross validated at sub-national levels to quantify the accuracy of the datasets. This study examined the utility of demographic mapping methods and how they have been used to address accuracy issues. It further cross validated WorldPop's estimates of pregnancies and live births using data (pregnancies and live births outcomes) that was collected as part of a project that was conducted in the regions of Maputo and Gaza provinces in Mozambique as the baseline data as the case study. The WorldPop dataset is one gridded global dataset that maps the population and demographic distributions of low income regions. Statistical analysis was used to determine the errors and performance of the WorldPop model and magnitude of errors at different administrative levels. Overall the results of this study showed that the Worldpop's pregnancies and live births datasets cannot yet be used without adjustments for the regions of Maputo and Gaza and there is need to improve the population allocation accuracy. Generally, the dataset exhibited a good variation modelling performance especially at higher administrative levels. The review of the demographic mapping methods also revealed better methods that are applicable that could be used in improving the accuracy and detail of global population distribution datasets, the WorldPop dataset included.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CDC - Centres for Disease Control

CLIP - Community Level Intervention for PreEclampsia

EEA - European Environment Agency

GIS - Geographical Information Systems

GPS - Global Positioning System

GPW - Gridded Population of the World

GRUMP - Global Rural Urban Mapping Project

LULC - Land Use/ Land Cover

MDG - Millennium Development Goals

MDSR - Maternal Death Surveillance and Response

NLCD - National Land Cover Dataset

PMC - PubMed Central

SDG - Sustainable Development Goals

SEDAC - Socioeconomic Data and Application Centre

UHC - Universal Health Coverage

UN - United Nations

UNEP - United Nations Environment Programme

UNFPA - United Nations Population Fund, formerly the United Nations Fund for Population Activities

WHO - World Health Organisation

CHAPTER 1: BACKGROUND AND OBJECTIVES

1.1 INTRODUCTION

This paper describes a scoping review performed to identify and review the methods used to produce population distribution datasets and the methods used to perform a cross validation of WorldPop's estimates for women of reproductive age populations, live births and pregnancies as a case study. The cross validation was performed using data, collected in the year 2015 as part of the baseline study for the Community Level Interventions for PreEclampsia (CLIP) trial, to enable the assessment of the accuracy of WorldPop's estimates of maternal and perinatal outcomes.

In many low and middle income counties, data on maternal and perinatal outcomes are not routinely or accurately collected, with national level estimates of these mostly only available from censuses that are conducted after ten year timelines at best (Tatem et al., 2007). Considering the importance of such data for measuring progress in improving maternal and perinatal health, and formulating relevant policy, new methods have been developed from different datasets including satellite imagery, and geo-demographic mathematical models to generate these data and make them widely available to policy makers.

The aim of this research is the review of the methods used to produce population distribution datasets, which are used in the development of demographic datasets, and the cross validation WorldPop's estimates for women of reproductive age populations, live births and pregnancies within regions of Gaza and Maputo provinces in Mozambique, using detailed census data collected on the ground. Data from the baseline study for the Community Level Interventions for PreEclampsia (CLIP) trial in the same area was used as the basis for this validation exercise. The results of this research will be useful in adjusting WorldPop births and pregnancies estimates where necessary, and will increase accuracy of these estimates while decreasing associated uncertainty.

1.2 BACKGROUND

1.2.1 GIS IN MATERNAL AND PERINATAL HEALTH

GIS is an essential tool in effectively targeting relevant population groups in need of health care interventions and achieving universal health coverage (UHC) provided the information is accurate and the data sufficiently exposes disparities within remote populations (Roth et al). The key to promoting UHC is the exposure of any hidden gaps in health service provision using sufficiently disaggregated data (Roth et al) that is reliable. Thematic mapping, spatial analyses and spatial modelling have been identified as the GIS methods that are valuable in policy discussions pertaining to maternal and perinatal health, relying greatly on volume, completeness, timeliness and accuracy of data (Ebener et al., 2015). Benefitting from these methods requires availability of and good quality data (Ebener et al., 2015) with the former having been met by the existence of freely available global population distribution datasets such as GPW, GRUMP, LandScan and Worldpop (combination of AfriPop, AsiaPop and AmeriPop).

The introduction of the millennium development goals (MDG) prompted the extensive use of national population datasets, especially in low income regions, to derive health metrics for applications in developing intervention programmes aimed at achieving these goals (Tatem et al., 2012). The justification for their utilisation is that they are standardized and considered to be of acceptable accuracy for national scale applications (Patterson et al., 2007). Such justification was acceptable since efforts made towards achieving the MDGs within the set deadline of 2015 (WHO, 2016a) focused on national level adjustments (Tatem et al., 2013; Alegana et al., 2015). This therefore prompted influential studies like Hay et al., 2009, Gething et al., 2010, Soares and Clements, 2011, Schur et al., 2011, etc. that validated these datasets and provided national level adjustments overlooking the heterogeneities at sub-national levels (Tatem et al., 2012). Due to the scale of their applications the majority of these national datasets have only provided global error estimates while acknowledging subnational variation (Patterson et al., 2007).

Existence of inequalities at sub-national levels has been identified despite the declination of maternal and new born mortality rates at national level, leading to the need to investigate such situations using geographical analysis (Ebener et al., 2015). Despite the fact that progress towards meeting MDG goals was measured at national levels (Tatem et al., 2013; Alegana et al., 2015), the spatial variations at subnational levels needed to be taken into account in devising policies for maternal and new born health care interventions.

Monitoring progress at national level overlooks the mostly affected remote areas (Ruktanonchai et al., 2016) meaning that although no improvement is evident in such regions, it will not be reflected since progress is measured at national level. The use of highest level of disaggregation to avoid masking of existing heterogeneity (Ebener et al., 2015) produces a sincere depiction of the progress in maternal and perinatal health care in developing countries. Accurate geographic analyses at sub-national levels are therefore of great necessity, requiring accurate geographic data. The need to assess accuracy of national population datasets at all sub-national levels arises from the current emphasis on sub-national monitoring of maternal and perinatal health progress that has been brought about by the new targets stated in the new Sustainable Development Goals (SDGs) recently announced by the United Nations (Ruktanonchai et al., 2016).

It is fundamental to accurately identify populations with the most need for health care interventions (Ruktanonchai et al., 2016) to effectively evaluate the performance of health care systems, thus providing evidence to support decision making concerning 1) planning for safer births and healthier new-borns and 2) resource allocation and improving access to maternal and perinatal health care (Stevens et al., 2015) as this is one of the main focuses in healthcare delivery (Ebener et al., 2015). Inaccurate identification of the populations in need of maternal health-care interventions has been one of the causes of the variations in the use of maternal health care (Say and Raine, 2007). The use of poor information in research and policy making leads to inefficient allocation of limited resources deterring the desired achievement of improved maternal and perinatal health quality. A true representation of the maternal and perinatal population distribution is therefore crucial in the successful implementation of interventions and it can only be achieved using accurate and highly disaggregated data. Emphasis is on accuracy and detail (Linard and Tatem, 2012) of the population distribution datasets as their applications have become more intensive and their implications more pronounced in the achievement of the new SDGs.

One of the factors that have been identified as affecting the improvement of quality of public health (maternal and perinatal health included) is demographic changes (Tatem et al., 2012). The rapid demographic changes influence the need for constant updating of the demographic datasets, giving rise to the need for up to date input data. The matter of temporal resolution of the input data is then bought into focus. Most developing countries lack data of high temporal resolution due to lack of resources required for the data collection. The available input data used for the establishment of the datasets come from multiple sources with errors of unknown magnitude which when combined with the errors arising from model assumptions contribute to the decrease of the accuracy of the datasets (Patterson et al., 2007).

Researches like Hay et al., (2005) have performed comparison of the accuracies of the databases to determine the database with the best accuracy. These national datasets have however become widespread and available for use such that comparison of their accuracies alone is not enough to justify their use. For such reasons the datasets require quantifying the quality of their data and examination of the impact of accuracy on the spatial model outputs, which are generated from using them.

The WorldPop dataset, this paper's case study, was created using land cover datasets and settlement extents together with the dasymetric modelling method (Linard et al., 2010) to produce a population distribution dataset with better accuracy and improved resolution of 100m, and has found many applications in the spatial analyses of maternal and perinatal health in developing countries (Tatem et al., 2014). However, information on data accuracy at specific locations within the datasets (Patterson et al., 2007) of population distributions should be made available.

1.2.1 IMPLICATIONS OF GLOBAL DATASETS' DATA IN MATERNAL AND PERINATAL STUDIES

Maternal and perinatal **s**tudies done for purposes of decision making, policy making and planning interventions, strictly need to be performed using accurate distribution maps produced from highly accurate input data. However due to the fact that overtime the demand for data used for these studies has increased and the geographic coverage of the areas of studies has also increased, mapping resources and the expertise needed to produce such maps have become a great challenge to satisfy the demand (Linard et al., 2012). Demand and coverage have been addressed by the availability of freely available national datasets. Little information is however provided to the users about data collection, processing and error distribution and the majority of national datasets only provide global error estimates without information on accuracy at subnational levels (Patterson et al., 2007).

Anderson (1976) proposed a 85 percent accurate standard baseline for national datasets but studies determined that the National Land Cover Dataset (NLCD) had an accuracy lower than the proposed standard baseline (Patterson et al., 2007). Such information should also be made available for the users of population distribution datasets to promote awareness of such uncertainties. The global data sets' utility is widespread and the implications of their data great. Users of such data therefore should know the associated errors of the datasets at subnational levels to reduce uncertainties associated with the data.

The dataset was created from multiple sources with errors of unknown magnitude and the cumulative error of the input data together with model assumptions introduced errors in the datasets (Patterson et al., 2007). Users therefore should be made aware of such information as the magnitude of error within the datasets, methods that were used to create the datasets and how these methods were used to deal with errors and improve accuracy of the datasets.

1.3 WORLDPOP'S POPULATION DATASETS – A CASE STUDY

As a source of data that is widely used for policy making and decision making, the implications of the WorldPop datasets are too great to simply ignore validating the accuracy of the data. Considering that its data is being used as substitute for performing household surveys at a large scale, to create unquestionably accurate demographic distribution maps, there should be conclusive evidence that the data is reliable for its intended purposes. With the limited resources, available for the health care intervention programs for these low-income regions, there is need for accurate input data for analyses done prior to making decisions to ensure targeting of the right population groups. Knowledge of the level of accuracy of the data they are using allows the researchers to factor in uncertainty brought about by the degree of accuracy of their input data. The assessment of the datasets brings the aspect of reliability to the attention of the users, thus cultivating a culture of always considering uncertainty of the WorldPop data. Quantifying the errors within the datasets, which will be part of the deliverables of this paper, encourages the users to also quantify the levels of uncertainties of the results obtained by using WorldPop data instead of instinctively trusting that the data will produce results that can lead to reliable conclusions simply because the datasets have a good resolution.

1.4 RESEARCH OBJECTIVES AND QUESTIONS

The objectives of this study are

1. To evaluate the utility of existing methods of demographic mapping for maternal and perinatal outcomes.
2. To cross validate WorldPop's estimates of women of live births and pregnancies for the regions of Maputo and Gaza provinces in Mozambique.

These objectives were met by addressing the following questions:

1. What are the main methods used for estimating the population distributions?
2. What is known concerning the levels of accuracy of these methods?
3. How accurate are the WorldPop's datasets in predicting maternal and perinatal outcomes?

1.5 DATA AND METHODS

The first objective was addressed through a scoping review method. The York methodology that was proposed by (Arksey and O'Malley, 2005) was used for this part of the project. The method consists of a five-part process including

1. Identifying the research question
2. Identifying relevant studies
3. Study selection
4. Charting the data
5. Collating, summarising and reporting the results

The second objective was addressed using data on maternal and perinatal outcomes in Mozambique. Neighbourhood level estimates of the outcomes under study have already been generated as part of the baseline work for the CLIP trial. For each of the neighbourhoods in the study, a corresponding value for each of the outcomes of interest was generated from the Worldpop data.

GPS coordinates of the inhabited areas are available from the project undertaken to determine the number of livebirths, and population distributions of women of the reproductive age. The dataset was assessed based on the actual extent of the populated areas and the available corresponding values for those neighbourhoods. A simple analysis using ordinary least squares regression was then performed between the WorldPop estimates of maternal and perinatal outcomes and the outcomes from the CLIP baseline census to determine the accuracy of the WorldPop dataset at different administrative levels.

1.6 DISSERTATION STRUCTURE

The next chapters highlight the methods used to address the objectives of the research, the findings of the research, analysis of the findings and the overall comments and conclusions derived from the entire research. Chapter 2 highlights the scoping review method used and the results of the scoping review, addressing the first objective of the research. Chapter 3 outlines the methods used to achieve the second objective which is to evaluate the accuracy of the WorldPop's dataset. Chapter 4 includes the results of the evaluation both descriptive and statistical. In chapter 5 the author analysed the results and highlighted their relevance in relation to the literature. The research was concluded in chapter 6, with the author summarising the findings and highlighting the limitations of the research and recommendations.

CHAPTER 2: SCOPING REVIEW

2.0 INTRODUCTION

This chapter addressed the first objective of this research by evaluating the utility of existing methods of population mapping for maternal and perinatal outcomes, through a scoping review which is a process of mapping the existing literature or evidence base (Arksey and O'Malley, 2005; Armstrong et al., 2011). The choice of this review method was influenced by the need to learn about the existing literature that has addressed the topic of demographic distribution mapping and identify any knowledge gaps from the literature.

2.1 SCOPING REVIEW METHOD

The York methodology that was proposed by (Arksey and O'Malley, 2005) was used for this part of the project. The method consists of a five-step process that involves firstly identifying the research question(s) to be addressed by the scoping review. The questions were used as a guideline in generating search terms that were then used in the search for academic literature. The first selection of articles was done by reading the abstracts and titles of identified literature and was followed by full text reading which led to the final selection of articles using themes. Key themes that the researcher was focusing on were generated, that would answer the research questions. These themes were used as headings in the charting of data. The results of charting were summarised and reported in the literature review section of the paper.

2.1.1 IDENTIFYING THE RESEARCH QUESTION

Identification of the research questions required the researcher to identify the questions that the scoping review would try to answer to achieve the first objective of the research. Two questions were addressed:

1. What are the main methods used for estimating the geographic distribution of populations?
2. What is known concerning the levels of accuracy of these methods?

2.1.2 IDENTIFYING RELEVANT STUDIES

Identification of relevant studies required the researcher to formulate keywords and a search strategy that would address these three questions. Keywords falling within the broad themes “methods (GIS, RS, etc.)” and “population & demography” were used to search for both academic articles and grey literature. A grey literature search was performed using Google and websites for key organisations like World Health Organisation (WHO), USAID, United Nations (UN), World Bank and Centres for Disease Control and Prevention (CDC) using the keywords “geographic distributions” and “maternal and perinatal”. The grey literature search was mainly focused on the application of the demographic distribution datasets in maternal and perinatal studies. The database used for the search for academic articles was PubMed Central.

2.1.2.1 GENERATION OF SEARCH TERMS

The main articles whose methods were used in the generation of the WorldPop dataset were searched on Google Scholar. Articles that cited these main articles were then scanned for keywords which were then extracted from either the abstracts or the keywords used in those articles. The same method was used on those articles to search for other relevant articles that cited them. The process was iterated until a set of keywords was generated that addressed the research questions formulated in the first step. This search resulted in 15 articles that were used to generate additional keywords.

Another method used for the generation of the search terms was the search for referenced articles within the main articles. Search terms were then extracted from the abstracts and keywords used. The selection of search terms was broken down into two themes. One was the “methods” used to generate the large-scale population datasets and the other was “population and demography”. The list of the search terms used is shown in Annexure A. The search on the PubMed Central database resulted in 652 articles. Articles that were a result of the search on Google Scholar were not part of the PMC articles as they were from different databases, therefore they were also added to the PMC articles, resulting in 667 academic articles.

2.1.3 STUDY SELECTION

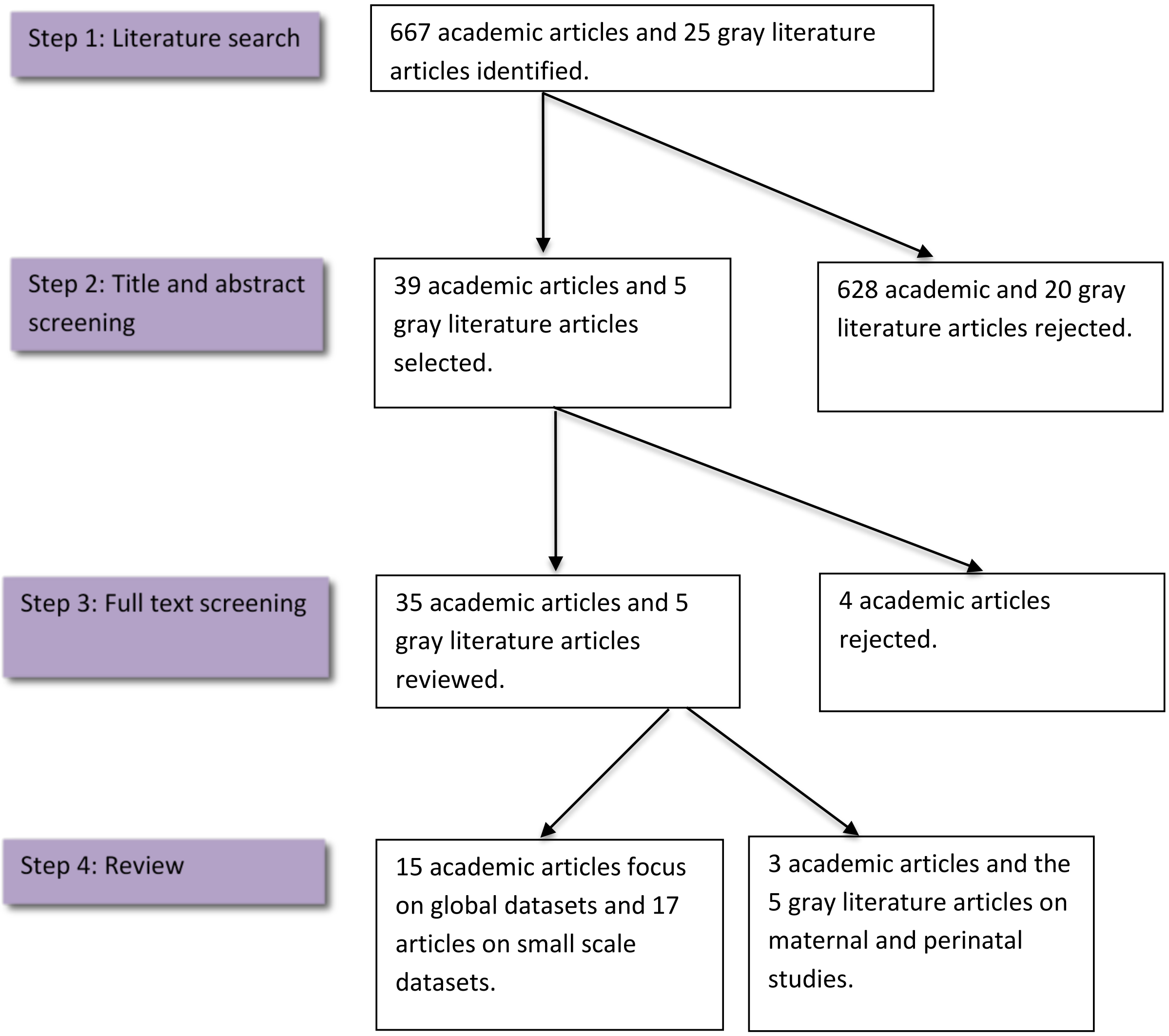
Study selection involved selection of the articles relevant to the study achieved by reading through the abstracts of the articles. Articles that focused on the methods used to generate the population datasets and those that focused on the use of these datasets in maternal and perinatal studies were selected. The selection of relevant articles from the PMC articles resulted in 24 articles to be reviewed, leading to a total of 39 academic articles being reviewed.

2.1.4 CHARTING THE DATA

The focus of the review was the methods that have been used in the creation of the demographic maps. In reviewing the articles, the researcher focused on the scale of the datasets, ancillary data used for disaggregating, redistribution modelling methods, spatial resolution of the input census data, accuracy assessments of the datasets and other statistical methods used for the generation of the datasets. The researcher paid attention to the way different authors argued about the methods they considered to produce datasets with better accuracy and detail.

2.1.5 COLLATING, SUMMARISING AND REPORTING THE RESULTS

This final stage of the scoping review was the final write up, where the researcher summarised the findings from the review of the articles, which constituted the literature review. The schematic representation of the processes involved from identifying the articles to the final selection of articles is shown in figure 2.1. The diagram only shows the results from the PubMed electronic database search and the google and website searches. Other articles were obtained from manually searching different search engines like GoogleScholar, ScienceDirect and Jurn.


Step 1: Literature search
667 academic articles and 25 gray literature articles identified.
Step 2: Title and abstract screening
39 academic articles and 5 gray literature articles selected.
628 academic and 20 gray literature articles rejected.
Step 3: Full text screening
35 academic articles and 5 gray literature articles reviewed.
4 academic articles rejected.
Step 4: Review
15 academic articles focus on global datasets and 17 articles on small scale datasets.
3 academic articles and the 5 gray literature articles on maternal and perinatal studies.

Source: Author
Figure 2.1: Flow chart of procedures involved from identification of articles to their review

2.2 RESULTS OF SCOPING REVIEW

2.2.0 WHAT IS KNOWN CONCERNING THE UTILITY OF POPULATION DISTRIBUTION MAPPING METHODS?

This section of the review results highlights the findings concerning the existing methods that have been used to develop population distribution maps.

2.2.1 GLOBAL POPULATION DATASETS

Global mapping efforts have been evident since the 1990s, where the producers have used a diversity of statistical methods and input data for the estimation of population on a global or regional scale in the production of gridded population maps. Such efforts include the Global Rural Urban Mapping Project (GRUMP), Gridded Population of the World (GPW), LandScan, United Nations Environment Programme (UNEP), European Environment Agency (EEA), Socioeconomic Data and Application Center (SEDAC) and WorldPop (formerly known as AfriPop and AsiaPop) (Dmowska and Stepinski, 2014; Dobson et al., 2000; Bhaduri et al., 2007; Linard et al., 2010; Nieves, 2016). The need for more accurate data sets has prompted the need to research methods that can be used to improve their accuracy. A summary of the methods and data used in the creation of these datasets is provided in a table (Table 2.1).

The combination of land cover data and settlements extents has been proven to improve the accuracy of global population datasets (Linard et al., 2010). The merits of using land cover datasets as ancillary data have been evident in global population datasets for low income regions where the land cover information has helped improve accuracy of population distribution data in countries where there is course resolution (Linard and Tatem, 2012; Alegana et al., 2015) census data of over a decade old and at provincial or district level (Linard et al., 2010; Linard and Tatem, 2012). Availability of high resolution (hi-res) population distribution datasets (defined as being at least 100m resolution datasets) at a large scale is however limited by the limited availability of hi-res land use/land cover (LULC) data sources at large spatial scales (Dmowska and Stepinski, 2014) as ancillary data. Ancillary data only increases accuracy of the modeling method if it is complete (Linard and Tatem, 2012) and has finer spatial detail (Tatem et al., 2007) than the input census data. As much as the accuracy of the disaggregated datasets depends on the resolution and age of the input census data, it is also dependent on the quality, significance and spatial (and temporal) resolution of the spatial covariate layers, created from the ancillary datasets, used to statistically aid the disaggregation (Patel et al., 2016).

Table 2.1: Global datasets, methods and data used

Dataset	Coverage	Resolution	Modelling technique	Ancillary data/boundary data	Availability
EEA	EU countries	100m	Dasymetric	CORINE LandCover 2000	Open-access
SEDAC	USA	1km & 250m	Areal weighting	US Census tracts	Open-access
LandScan USA	USA	90m	Multi-dimensional dasymetric	National Land Cover Dataset	Commercial
WorldPop	Africa, Asia, Central and South America	100m	Dasymetric	Africa: GlobeCover Asia & Americas: MDA GeoCover	Open-access
GPW	Global	5km	GPW1: Pycnophylactic GPW2: Areal GPW3: Areal	Census, water bodies (for masking)	Open-access
UNEP	Global	5km	Smart interpolation	Census, populated points, roads	Open-access
LandScan	Global	1km	Smart interpolation	Census, land cover, elevation, slope, roads, populated areas/ points	Commercial
GRUMP	Global	1km	Dasymetric	Census, populated areas, water bodies (for masking)	Open-access

2.2.2 REVIEW OF THE WORLDPOP PROJECT METHODS – A CASE STUDY

The WorldPop project is one such data source that “provides an open access archive of spatial demographic datasets for Central and South America, Africa and Asia to support development, disaster response and health applications”. Of relevance to maternal and perinatal health, WorldPop produces spatial datasets at 100m resolution for estimates of populations of women of reproductive age, live births and pregnancies projected through 2035, using available data sources including census, Demographic Health Surveys and UN data, satellite imagery data, and estimates on stillbirths, abortions, and miscarriages (Tatem et al., 2014). These data are increasingly being used as part of global health programs and policy formulation; for example a recent publication in nature microbiology identified populations of reproductive age women at risk of adverse fetal outcomes from the pending Zika virus epidemic in the Americas (Alex Perkins et al., 2016).

The WorldPop dataset uses census data, available highest level administrative unit boundaries and official population estimates for each country (Dmowska and Stepinski, 2014) as input data sources and the dasymetric method is used to disaggregate census data using land cover data from GlobeCover for Africa (Linard et al., 2012; Tatem et al., 2007) and land cover data from MDA GeoCover for Asia and Americas (Gaughan et al., 2013) and settlement extents data (Linard et al., 2012; Tatem et al., 2007; Gaughan et al., 2013; Stevens et al., 2015). The comparison of four large scale, medium resolution land cover datasets (AVRR-derived, MODIS-derived, GLC 2000 and Globcover) yielded the results that led to the selection of the Globcover dataset as it yielded the most accurate population distribution models over the other three (Linard et al., 2010). This led to its use as the land cover data source for mapping the African continent as part of the WorldPop project.

The very course temporal and spatial resolution of census data, or lack of strong input data, for most low income countries, mapped by the WorldPop dataset, are a great limitation in the development of the dataset (Linard et al., 2010). To address this limitation models developed form countries with high spatial and temporal census data resolution, like Kenya, are partially parameterized on neighbouring countries (Stevens et al., 2015) which are spatially proximate and environmentally similar (Linard et al., 2010) despite the fact that this further conceals the already non-intuitive relationships between population density and the supporting covariates (Nieves, 2016).

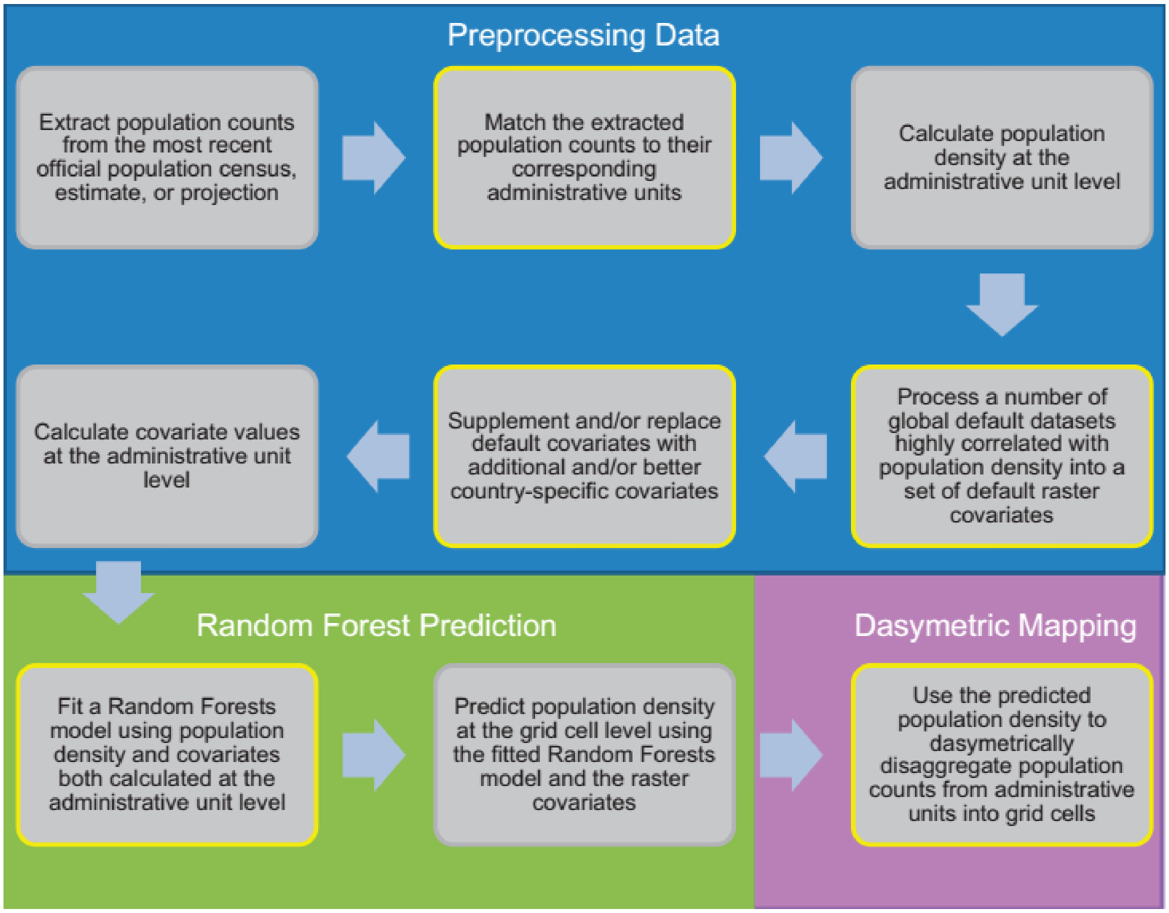
Conversion of population distribution datasets to gridded estimates of births and pregnancies is achieved through the integration of " household survey data, UN statistics and other sources on growth rates, age specific fertility rates, live births, stillbirths and abortions" (Tatem et al., 2014).

The emphasis is therefore on the methods used to create the population distribution datasets as their level of accuracy determines the accuracy of the gridded estimates of births and pregnancies.

2.2.1.0 WORLDPOP RANDOM FOREST-BASED DASYMETRIC POPULATION MAPPING

WorldPop uses a Random Forest regression model and dasymetric mapping methods in a

three step method to produce the disaggregated population distribution dataset. These steps include (1) covariate selection for the Random Forest model, (2) fitting of the Random Forest model and creation of a population density weighting layer from the created Random Forest model and (3) dasymetric redistribution of population counts (Stevens et al., 2015).


Preprocessing Data
Extract population counts from the most recent official population census, estimate, or projection
Match the extracted population counts to their corresponding administrative units
Calculate population density at the administrative unit level
Process a number of global default datasets highly correlated with population density into a set of default raster covariates
Supplement and/or replace default covariates with additional and/or better country-specific covariates
Calculate covariate values at the administrative unit level
Random Forest Prediction
Fit a Random Forests model using population density and covariates both calculated at the administrative unit level
Predict population density at the grid cell level using the fitted Random Forests model and the raster covariates
Dasymetric Mapping
Use the predicted population density to dasymetrically disaggregate population counts from administrative units into grid cells

Source: Gaughan, A.E., Stevens, F.R., Huang, Z., Nieves, J.J., Sorichetta, A., Lai, S., Ye, X., Linard, C., Hornby, G.M., Hay, S.I., Yu, H., Tatem, A.J., 2016. Spatiotemporal patterns of population in mainland China, 1990 to 2010. Sci. Data 3, 160005. doi:10.1038/sdata.2016.5

Figure 2.2: Flow diagram of the WorldPop approach to mapping population

2.2.3 NATIONAL AND REGIONAL SCALE POPULATION DATASETS

Global datasets have the drawbacks of being course and general (Jia et al., 2014), leading to the focus of the small-scale datasets which is to address the limitation of detail and accuracy in global datasets. Studies have tried to focus on regions which are part of the mapped spatial extents within the existing global datasets. These studies have made emphasis on investigating the effects of using different ancillary data sources, and in some cases methods, to enhance the existing datasets. Another emphasis has been to compare the results they obtain using their suggested methods and data sources with those they produced using the same methods and data sources as those used by the producers of global datasets. Datasets created for the purpose of enhancing existing datasets include Indonesia (WorldPop) (Patel et al., 2015), Kenya (WorldPop) (Deleu et al., 2015) and the United States of America (SEDAC) (Dmowska and Stepinski, 2014). Other studies like (Lung et al., 2013), (Jia et al., 2014), (Cockx and Canters, 2015) and (Douglass et al., 2015) focused on comparing their methods with those used to produce global datasets. Their aim was to determine whether the methods and data used in global mapping efforts are adequate for producing more detailed and accurate datasets or use of alternative and refined methods and data produced better results. These studies have proven that the detail and accuracy of a dataset can be improved by ancillary data with finer detail. A summary of the data and methods used in the production of these datasets is presented in table 2.2.

2.2.4 EFFECTS OF CHOICE OF ANCILLARY DATA SOURCES

Some ancillary data sets selected for disaggregating populations represent phenomena known to be correlated with population densities, for instance, it has been identified that humans tend to modify their environment, specifically land cover, in ways that differentiate it from the surrounding landscape (Nieves, 2016) . Satellites have been the most commonly used source of ancillary data in the form of land cover (Lung et al., 2013) and land use data used for estimation of population densities because of the high correlation between land use/land cover (LULC) category and population density (Dmowska and Stepinski, 2014; Yang et al., 2009). Some remotely sensed data sources used for large scale demographic maps however have resolutions that are too low for obtaining accurate disaggregated data especially for urban areas which are highly heterogeneous (Cockx and Canters, 2015).

The limitation of using remotely sensed data (whether high or low resolution) is that it cannot be reliably derived by any known algorithm (Jia et al., 2014) due to the assignment of weights to the LULC classes being based on heuristic rules and assumptions without a solid evidence base for such rules (Lung et al., 2013). Another limitation of using land cover data, especially in heterogeneous urban areas, is the overestimation of population densities in certain land cover classes like “developed, open space”, due to the category being intermingled with other urban categories having high population density (Dmowska and Stepinski, 2014). However, Lung et al (2013) further explore the possibility of the opportunities of the very high satellite imagery currently having not been fully explored. This notion leaves room for the development of more detailed datasets with better accuracy using remotely sensed data.

Other types of remotely sensed data like spectral and/or textural metrics or demographic information and distance-to-services metrics are lower-resolution ancillary data that have also been applied, with the results obtained in these studies suffering from low accuracies (Cockx and Canters, 2015). This has been proven not to be the case with high resolution remotely sensed data sources as they can provide such data types in greater detail for disaggregating the census data although such data sources have low coverage making their application in large scale mapping limited or impossible (Cockx and Canters, 2015). This insinuates that use of these data sources at higher and more detailed resolutions, as ancillary datasets, increases their potential of better performance in producing datasets with better accuracy.

The use of high resolution ortho-rectified RapidEye archive data for settlement extents to enhance the Afripop dataset for the border region comprising South-Africa, Swaziland and Mozambique proved to improve the accuracy of the dataset showing more detail for the study areas (Deleu et al., 2015). The results showed that this method had a high potential of being replicated for the other countries to allow improvement especially in the detail of the dataset. Comparison of the Afripop data set and the newly created data set for two cities Matola and Jozini are shown in figure 2.2. This due to the utility of the data source being independent of the influence of any population characteristic. Another data source that has proven to be a reliable source of ancillary data but unlike the RapidEye archive data does not possess a guaranteed potential to be replicated are the geolocated tweets. This has been shown to be a powerful ancillary data source for regions with highly active tweets like in Indonesia (Patel et al., 2016). Integration of geotweets data into the methods used in the production of the AsiaPop dataset proved to improve the accuracy of the dataset for the study area Indonesia but unlike the high resolution settlements data that was used by Deleu et al., (2015) to enhance the AfriPop dataset, this method of using geotweets does not have the potential to be replicated by any other country (Patel et al., 2016). The strength of its application is in the density of geotweets in the whole country (Patel et al., 2016), i.e. the higher the density of active twitter users the greater the potential of the use of this method.

Impervious surfaces (roads, sidewalks, driveways and parking lots, that are covered by impenetrable materials), have been proven to perform equally well as or better than land use data as a source for disaggregating population data (Cockx and Canters, 2015). Another source of ancillary data that has been confirmed to be an appropriate alternative ancillary data set instead of land cover that may provide the necessary increased level of spatial detail in population counts is parcel data as it has a specific orientation towards population density for a given point in space particularly in rural area (Jia et al., 2014). The comparison between parcel derived outputs and land cover derived outputs shows that parcel derived outputs produce more accurate distributions than land cover derived outputs (Jia et al., 2014). The output datasets produced form using parcel data and land cover data are shown in figure 2.4. As seen in the figure, parcel data produces a more detailed population distribution dataset. Another parcel related data source that has been used in other studies is the residential address points data where in a comparison between the use of residential address points with dasymetric mapping techniques based on land use, impervious surface cover, light emission and road density studies have shown that methods using address points outperform more traditional dasymetric mapping techniques (Cockx and Canters, 2015). Despite parcel related data being used in a few studies, no efforts have been made to explore the relationship between parcel type and population density (Jia et al., 2014).

Integration of the parcel data and land cover data increases the overall accuracy of gridded population surfaces (Jia and Gaughan, 2016). Combination of the strengths of these two data sources proved to increase accuracy, verifying that accurate datasets are produced by taking advantage of the strengths of different types of ancillary data instead of choosing one over the other. Of course, the choice of ancillary data sets is critical, as the datasets need to have an acceptable correlation for them to complement each other.

Table 2.2: National and sub-national datasets methods and data

Author(s)	Date	Spatial Coverage	Resolution	Input spatial data resolution	Population distribution modelling method	Ancillary data
Lung et al	2013	Rural Kenya	1km	-	pycnophylactic & smart interpolation	QuickBird (0.6m panchromatic & 2.4m multispectral bands), roads, rivers/streams, schools, markets, slope gradient
Jia et al	2014	Alachua county(USA)	30m	census tracts	dasymetric	2010 parcel data
Cockx et al	2015	urban Belgium regions	1km	-	dasymetric	address point information
Deleu et al	2015	Kenya	100m	Sublocation unit boundaries	dasymetric	RapidEye archive data (settlement extents) & GlobeCover (land cover)
Jia et al	2016	Alachua county (USA)	30m	census tracts	dasymetric	land cover and parcel data
Dmowska et al	2014	USA	90m	SEDAC 1km & 250m grids	dasymetric	NLCD 2001 and 250m SEDAC-MSA
Douglass et al	2015	Lombardy region (Northern Italy)	1km	-	dasymetric	telecommunications data
Patel et al	2016	Indonesia	100m	-	dasymetric	Geolocated tweets
Deville et al	2014	Portugal & France	1km	ADM-5 “Freguesias” in Portugal and “Communes ” in France)	dasymetric	mobile phone geolocation data
Mennis	2003	Pennsylvania	100m	census block	dasymetric and areal	Landsat TM imagery

Use of mobile phone geolocation data to disaggregate census data has been proven to improve the accuracy of population densities as it captures the dynamic nature of populations (Deville et al., 2014) providing “real time measure of populations” (Douglass et al., 2015). Mobile phone communications are located by identifying the geographic coordinates of their transmitting towers and the associated cells, making the network-based positioning method simple to implement. However, like geo-located tweets, its accuracy is directly dependent on the network structure, thus the higher the density of the towers, the higher the precision of the mobile phone communication geo-location (Deville et al., 2014). In a comparison between the performance of datasets created using the mobile phone and remote sensing methods, the results showed that the remote sensing method produced predictions with a higher precision but less accuracy, with an over-estimation of population densities in low-density areas and an under-estimation of population densities in high-density areas (Deville et al., 2014). Although at a global level, the remote sensing method was proven to be more precise than the mobile phone method the mobile phone method exhibited an overall better performance as it was slightly more accurate than the remote sensing method (Deville et al., 2014). Figure 2.3 shows the comparison of the datasets produced by the mobile data method and remote sensing method (land use and land cover data).

While Deville et al., (2014) focus on the use of mobile data in application to census redistribution, the potential of telecommunications data has been explored the its strength in other applications like on predicting inter-census period population using models trained on known census data (Douglass et al., 2015).

2.2.5 SPATIAL INTERPOLATION ALGORITHMS

For purposes of confidentiality and privacy concerns census data has for the past years always been collected at household level, then aggregated and presented in choropleth form (Jia et al., 2014; Dmowska and Stepinski, 2014). Such representation of population distributions assumes homogenous distributions of populations within administrative boundaries. Population redistribution models have been developed to disaggregate these census data into uniform grids that are independent of the administrative boundaries, leading to the creation of demographic maps that can be used for spatial analysis and modelling that is independent of the political boundaries created within different countries. Disaggregation, spatial decomposition (Bhaduri et al., 2007) or downscaling (Gallego, 2010) is the transformation of unit-based data into raster-based data having cell size smaller than a majority of areal units (Dmowska and Stepinski, 2014). Spatial interpolation algorithms are the methods used to incorporate ancillary information into population distribution modelling (Lung et al., 2013). Four spatial interpolation algorithms have been used by large scale population datasets to reallocate census data within administrative units into continuous uniform grids (Linard and Tatem, 2012).

2.2.5.1 AREAL WEIGHTING

Areal weighting was used in the construction of the GPW2 and GPW3 datasets (Hay et al., 2005). The assumption is that population is uniformly distributed within the same administrative unit thus the value in each cell becomes the population of the administrative unit divided by the number of cells within the unit (Linard and Tatem, 2012). Identified drawbacks of areal weighting are its assumption of a homogeneous population distribution within an administrative unit and the introduction of the problem of artefacts at administrative boundaries (Lung et al., 2013). The areal weighting method provides the most accurate results when very fine-resolution census data are available (like in the case of Kenya where the spatial resolution of the census data was at sub-location level) and the use of land cover data at such high spatial resolutions in population distribution modelling does not inevitably improve the simple areal weighting method (Linard et al., 2010) due to the low resolution of the satellite derived land cover data used.

2.2.5.2 PYCNOPHYLACTIC INTERPOLATION

Pycnophylactic interpolation was used in the construction of the GPW1 database (Hay et al., 2005). This method starts with the use of areal weighting then the population values are smoothed using weighted mean of the nearest neighbouring units while preserving the original total populations of the administrative units (Linard and Tatem, 2012). This modelling technique addresses the problem of artefacts (ie unrealistic changes in population densities) by smoothing the population distribution at the boundaries (Lung et al., 2013).

2.2.5.3 DASYMETRIC MODELLING

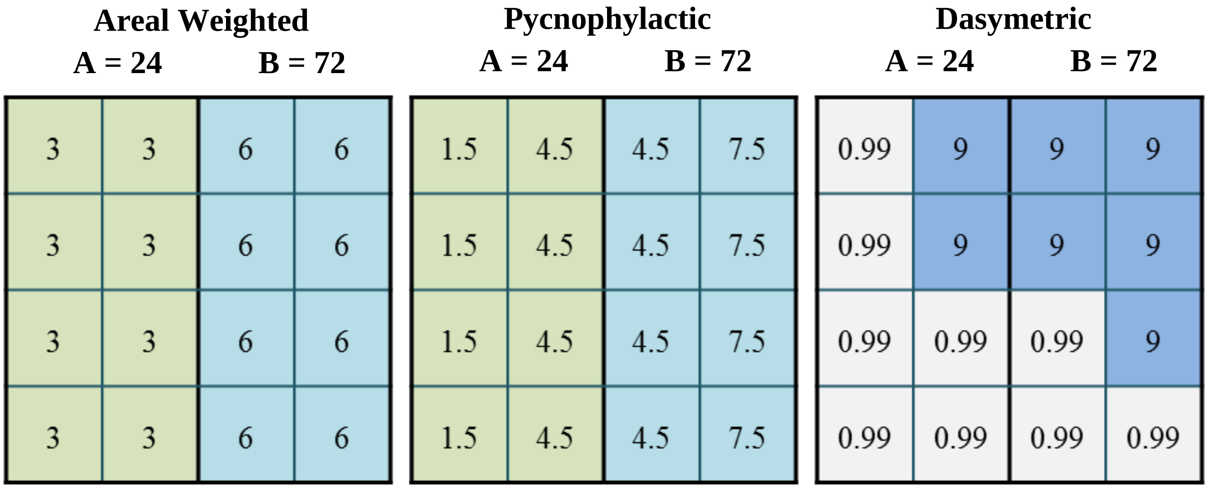
Dasymetric modelling embroils use of ancillary data - often including satellite derived land cover data and other data like slope and roads - to redistribute populations within administrative units according to weightings attributed to different land cover classes (Linard and Tatem, 2012). The WorldPop project uses this approach whereas the GRUMP database incorporated this method with the GPW methods (Linard and Tatem, 2012). Dasymetric methods have been applied mostly at local and regional scales (Dmowska and Stepinski, 2014). Overestimation of areas with low population densities and the underestimation of areas with high population densities have been the drawbacks of dasymetric methods frequently experienced (Cockx and Canters, 2015). Various techniques have been used to increase the accuracy and detail of dasymetric modelling approaches (Maantay et al., 2007; Mennis, 2009; Wu, Qiu, & Wang, 2005), of which one well-cited approach is the Heuristic Sampling Method (HSM) (Mennis, 2003; Jia et al., 2014).

2.2.5.4 SMART INTERPOLATION

Smart interpolation, which was used in the construction of the UNEP and LandScan datasets (Hay et al., 2005), involves modelling population distributions at a finer scale using satellite and other ancillary data (referred to as proximity factors (Lung et al., 2013)) like road network, slope, terrain, land cover and night-time lights to determine the probability of existence of populations in cells (Linard and Tatem, 2012). The influence of these factors on populations is used to assign single coefficients to them from which an overall weight is calculated for each grid cell and used for population distribution (Lung et al., 2013). The overall weights assigned to each cell are derived from the proximity of the factors to the cell which determines their influence on the cells.

Uncertainties and errors in these large scale population datasets arise mainly from input data, temporal projections and the modelling method used together with the inaccurate spatial positioning of administrative boundaries (Linard and Tatem, 2012). Inaccurate positioning of administrative boundaries increases uncertainties in spatial analyses as an overlay of such datasets with a boundary dataset with good positional accuracy yields less accurate results with incorporated errors from mismatching boundaries. Effects of such uncertainties become highly evident in higher level administrative units (Tatem et al., 2013). These effects amplify the strengths of dasymetric and smart interpolation methods over areal weighting and pycnophylactic interpolaton methods as the former are mostly independent of boundary data and the latter are dependent on boundary data.

The population of two administrative units, A and B (total population in A = 24 and B = 72) are redistributed using different spatial interpolation approaches (areal weighting, pycnophylactic and dasymetric). In the dasymetric method, a higher weight was attributed to the blue area.



Areal Weighted
A = 24 B = 72

3	3	6	6
3	3	6	6
3	3	6	6
3	3	6	6

Pycnophylactic
A = 24 B = 72

1.5	4.5	4.5	7.5
1.5	4.5	4.5	7.5
1.5	4.5	4.5	7.5
1.5	4.5	4.5	7.5

Dasymetric
A = 24 B = 72

0.99	9	9	9
0.99	9	9	9
0.99	0.99	0.99	9
0.99	0.99	0.99	0.99

Source: Author
Illustration adopted from Linard, C., Tatem, A.J., 2012. Large-scale spatial population databases in infectious disease research. Int. J. Health Geogr. 11, 7. doi:10.1186/1476-072X-11-7
Figure 2.3: Schematic illustrations of spatial interpolation methods.

Smart interpolation methods and dasymetric modelling evidently produce more detailed distribution models, proving that the use of satellite derived data together with other ancillary data improves the level of detail and accuracy in population distribution models. This was illustrated by (Linard and Tatem, 2012) when they compared population density distributions of two regions, in Kenya (with detailed spatial data) and Angola (with course resolution spatial data), from the different datasets produced using the four different modeling techniques.

Accuracy of the population distribution models is improved by the use of high resolution census data (that is very recent with a finer spatial resolution) and satellite derived data together with highly detailed ancillary data (Linard and Tatem, 2012). Limitations of spatial interpolation methods are the abrupt changes in population density and areas of no population which haven't been accurately reproduced by any of the modelling techniques (Lung et al., 2013). Lung and colleagues (2013) combined two modelling techniques in their efforts to produce a dataset that tried to address these limitations as much as possible in rural Kenya by combining the pycnophylactic and smart interpolation techniques. These methods did not prove to address these limitations but the effects of using the pycnophylactic technique to smooth the population densities at the boundaries was revealed. Although both methods preserved the aggregated population values, as the difference between their aggregated population values and those of the choropleth map were very little, there was an evident difference in population densities at the boundaries, produced by the two modelling methods.

2.2.6 ACCURACY OF THE METHODS

Currently there is no standard method that has been set to quantify the accuracy of the gridded population datasets. This is partly because the validation of these datasets in their entirety is not possible due to the lack of field data that can be used as a baseline (Linard and Tatem, 2012) for assessment. All the studies involved in this research have used error statistics to compare the performance of their datasets to the existing datasets or to compare the performance of different methods and data used in producing the same dataset. Statistics that have been mostly used for the assessment of accuracy include the root mean square error (RMSE), normalised root mean square error (NRMSE / %RMSE), mean absolute error (MAE), coefficient of variance (CV), coefficient of correlation (CC), standard error (SE) (Tatem et al., 2007) and the error matrix applying stratified random sampling (Lung et al., 2013; Deleu et al., 2015). Table 2.3 gives a summary of the error statistics obtained from the different studies.

2.2.7 APPLICATION OF THE DATA IN MATERNAL AND PERINATAL HEALTH STUDIES

The newly established SDGs are more ambitious than the MDGs as they have advanced form reducing maternal and perinatal mortality at national levels to reaching the most disadvantaged people everywhere requiring high quality data (WHO, 2016b). The National Centre for Chronic Disease and Prevention’s division for Reproductive Health has partnered with WHO, UNFPA, USAID and intergovernmental organisations to develop and implement a Maternal Death Surveillance and Response (MDSR) system with the aim to create a platform for interactive choropleth mapping of maternal and perinatal mortality indicators allowing for timely subnational level mapping (CDC, 2016). This platform incorporates the use of global maps and real time geographic maternal and perinatal indicator maps developed using uploads of maternal death counts and the projected women of reproductive age and live births counts by users from different countries (CDC, 2016). This system will play a big role in promoting better programmatic planning as it facilitates subnational level mapping which is crucial in determining effectiveness of lower level interventions.

Table 2.3: Summary of error statistics for gridded datasets

Author(s)	**Statistical method(s) used**	**Findings**
Lung et al. (2013)	Stratified random sampling	Producer's accuracy = 82%, User's accuracy = 95%
Patel et al. (2016)	MAE, RMSE, %RMSE	Using geotweets data enhanced the AsiaPop dataset by: RMSE = 70.15, %RMSE = 2.29%, MAE = 3.28 at admin level 4
Jia et al. (2014)	RMSE & CV	Parcel data produced RMSE = 63.96, CV = 1.36 and land cover data produced RMSE = 73.26, CV = 1.36. t-Test proved the difference is significant
Douglass et al. (2015)	RMSE	RMSE for: Land cover only = 232, Telecommunications data only = 227, Combined = 200
Jia et al. (2016)	RMSE & CV	RMSE for best combination of parcel data and land cover data = 59.22, CV = 1.08
Deleu et al. (2015)	RMSE, MAE, CC, Stratified Random Sampling	AfriPop MAE = 1063.36, RMSE = 1729.59, CC = 0.48 against Enhanced AfriPop MAE = 607.99, RMSE = 974.16, CC = 0.75
Gaughan et al. (2013)	RMSE, MAE, %RMSE	AsiaPop: Cambodia; RMSE = 3834.51, %RMSE = 46.40 MAE = 2494.32 Vietnam; RMSE = 4943.31, %RMSE = 70.13, MAE = 3007.04
Tatem et al. (2007)	RMSE & Standard Error	Afripop: EA1; RMSE = 592.1475, SE = 530.64 EA2; RMSE = 1097.754, SE = 971.51 EA3; RMSE = 574.1875, SE = 509.7
Alegana et al. (2016)	MAE, RMSE & CC	1km by 1km WorldPop grids: MAE = 0.0326, RMSE = 0.0426, CC = 0.6346
Deville et al. (2014)	RMSE	RMSE (Mobile Phone data) = 796; RMSE Remotely sensed data = 850
Stevens et al. (2015)	RMSE, MAE, %RMSE	**Cambodia**: AsiaPop %RMSE = 46.40 GRUMP %RMSE = 81.89, GPW %RMSE = 82.22 **Vietnam**: AsiaPop %RMSE = 70.13 GRUMP %RMSE = 92.56 3771.64 GPW %RMSE = 100.47 **Kenya**: AsiaPop %RMSE = 120.28 GRUMP %RMSE = 145.35 GPW %RMSE = 146.11

Neal et al., (2016) have used the WorldPop data together with Demographic Health Surveys data to determine the geographic distribution of adolescent first births in Tanzania, Kenya and Uganda. The aim of this study was to inform policy makers at district and national level of the heterogeneous distribution of adolescent pregnancies, with the focus being the “hot spots”, to assist in their policy making aimed at reducing adolescent pregnancies which have been proven to affect the health of adolescent mothers and their first bon children (Neal et al., 2016).

Perkins et al., (2016) created a model that projected the population of child bearing women that would be affected by the Zika virus in the highly affected Americas. Their aim was to improve the estimates made on the affected number of child bearing women and to ascertain the portion of children that were at high risks of being affected by the virus (Perkins et al., 2016). This study demonstrated the power of demographic maps in assisting targeted interventions in reducing maternal and perinatal deaths. Ruktanonchai et al., (2016) examined the sub-national heterogeneities in accessing maternal and perinatal health care before, during and after delivery and adjusted the 2015 live births estimates to assess the probabilities of accessing health care. In these studies, focus has been on subnational level assessment, highlighting the progression of maternal and perinatal studies and the growing relevance of high quality and detailed demographic data in maternal and perinatal health studies.

2.3 RESEARCH KNOWLEDGE GAPS

In this chapter the researcher has reviewed the different methods and data that have been used to develop gridded population datasets. In their studies the authors have mainly focused on comparing their results with results obtained from using remotely sensed land cover data as a source of ancillary data. This decision has been influenced using the land cover/ land use data in most global population datasets. Their results have shown that use of alternative ancillary data or the combination of such data with land cover/ land use data produces more detailed and more accurate results than using remotely sensed land cover/ land use data. None of these studies however have compared performance of these ancillary datasets against each other to determine which ancillary dataset performs the best and under what conditions they perform better over other datasets. It has been shown that the heterogeneity of land cover influences the difference in the performance of different ancillary datasets in producing accurate models. This means that the same method also performs differently within the dataset as we move from urban to rural settings. One ancillary dataset can perform better than the other in an urban setting but the reverse might be true in a rural setting. Currently methods are still being explored to finally come to a decision on which ancillary datasets (or combination of datasets) produces the best results.

Focus has been on validating population datasets despite the existence demographic datasets within the global datasets. Methods that have been used have been focused on population distributions. Studies have no yet revealed the importance of demographic datasets as relevant data for describing population distribution patterns.

The validation of datasets has been mainly relative, with the focus being on relative accuracy and not absolute accuracy of datasets. Spatial scale of validation also differs from one author to another. This means that comparison of performance of methods used by different authors cannot be done since modelling methods and ancillary data used perform differently at different spatial scales (Deleu et al., 2015). Performance of each dataset at different spatial scales has not been explored. Datasets have been compared against each other at single administrative unit levels. It has not been revealed and proven that certain methods perform better than the others at all spatial scales. As Deleu et al., (2015) suggested that areal weighting performs better than dasymetric modelling when the census data exists at very fine spatial detail, such findings have not been explored by several studies.

2.4 CONCLUSION

Existing literature has magnified the significance of the choice of ancillary dataset and spatial interpolation methods used in the production of population distribution maps. This chapter highlighted the methods currently being used in the production of these maps and the identified questions that have either not been addressed or fully addressed by the existing literature in relation to the objectives of this research. The WorldPop dataset was used as a case study in evaluating the accuracy of the methods used creating the global population datasets. The methods used and the results obtained from the evaluation are highlighted in the next chapters.

CHAPTER 3: METHODS FOR EVALUATING THE WORLDPOP DATASET

3.1 INTRODUCTION

As part of addressing the research questions that were posed earlier in Chapter 1, a regression analysis was completed using observed field data (CLIP data) and modelled data (WorldPop data). Vector grids were created from the 100m resolution WorldPop raster grids and the WorldPop and CLIP pregnancies and live births values calculated for each grid. The resulting table had fields with six administrative unit level names (neighbourhood code, new neighbourhood, cluster, locality, administrative post and district), WorldPop's live births and pregnancies estimates per grid and CLIP live births and pregnancies outcomes per grid. Each record in the attribute table represents a grid and its attributes. A grid cell level analysis was done using RMSE and CV, and these values were compared for both the pregnancies dataset and the live births dataset. A Welch's t-Test was done to determine the significance of the mean residuals for pregnancies and live births datasets. The results were also aggregated at different administrative unit levels and for each administrative level RMSE, CV and %RMSE values were calculated for analysis of the WorldPop dataset. As part of the analysis, charts were used to represent the trends exhibited by the modelled WorldPop estimates relative to the observed CLIP outcomes.

3.2 STUDY AREA

Maputo and Gaza are provinces located in Mozambique. According to the 2007 census the population of Mozambique was 20,252,223 (City Population, 2016) with a projected population of 28,751,000 in the year 2016 and an average annual population growth rate of 2.8% (UNdata, 2016). According to UNdata, (2016) 2010-2015 statistics, Mozambique has a 32.2 % urban population with an estimated growth rate of 3.3%, a fertility rate (live births per woman) of 5.5% and an infant mortality rate of 64/1000 live births. UNICEF, (2011) has reported a decrease of maternal mortality in Mozambique from 1,000 maternal deaths/100,000 live births during the early 1990s to an improved 408 maternal deaths/100,000 live births in 2003 (UNdata, 2016). The 2007 census showed that the province of Maputo had a population of 1,782,400 (City Population, 2016) with a projection of 1,709,058 by the year 2015 (GeoHive, 2016) and the province of Gaza had a population of 1,442,100 (City Population, 2016) projected at 1,416,810 (GeoHive, 2016). Maputo covers an area of 22,693 sqkm whilst Gaza covers an area of 75,334 sqkm (GeoHive, 2016).

3.3 DATA AND SOURCES

Data on pregnancies and livebirths in the study area (Maputo and Gaza provinces in Mozambique) were downloaded from the WorldPop website (www.worldpop.org.uk). GPS coordinates of the inhabited areas were available from the project undertaken to determine the number of live births, and population distributions of women of the reproductive age in the Gaza and Maputo provinces. Neighbourhood level estimates of the outcomes under study had already been generated as part of the baseline work for the CLIP trial. The data of interest available in a spreadsheet were the administrative unit names (district, locality, neighbourhood code, cluster, administrative post and new neighbourhood) and the aggregated cluster data of pregnancies and live births corresponding to each administrative unit. Each GPS coordinate represented a household, hence the spatial resolution of the baseline data used was at household level, the highest level of spatial resolution for census data. The values of live births and pregnancies were calculated for each GPS coordinate using data that was aggregated at neighbourhood level. The formula is shown below.

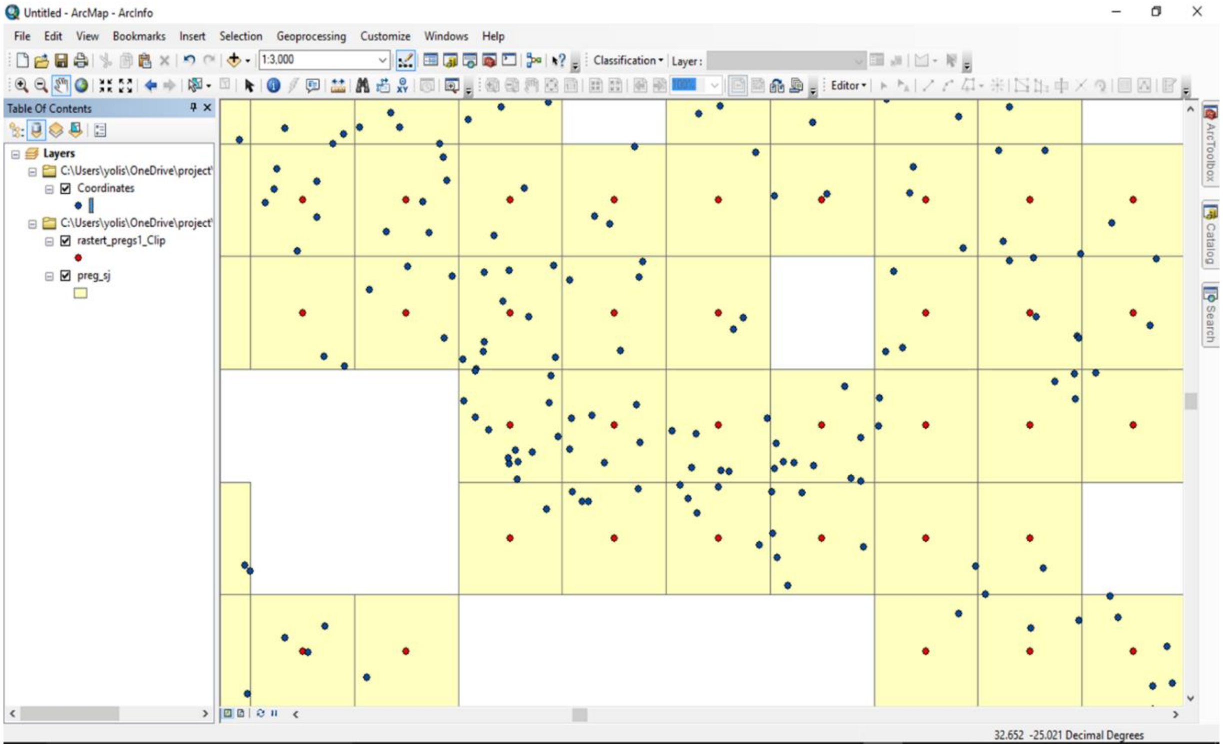
$$Value\ per\ GPS\ point = \frac{Population\ of\ neighbourhood}{Number\ of\ households\ (GPS\ points)\ neighbourhood}$$

This was done by dividing the number of pregnancies and live births at each neighbourhood by the number of GPS points (representing households) within that neighbourhood. The outcome of this process was a table with average CLIP pregnancies and live births per household with the corresponding neighbourhood code and other administrative unit names. The table was then joined with the GPS coordinates layer's attribute table on ArcMap. This process is done using the "Joins and Relates" tool that is found when you left click the layer whose attribute table you want to join with another table. This step is possible if both the attribute table and the table to be joined have a common field name that will be used to join corresponding attributes. For this research the common filed name that was used was the "neighbourhood code". This meant that GPS points with the same neighbourhood code were assigned the same values for pregnancies and live births.

Creation of 100m vector grids and calculation of Worlpop and CLIP values

The floating raster layers for WorldPop's live births and pregnancies were converted to integer raster layers then to vector points. The vector points layer was then used to extract the raster values from the two datasets using the "extract values to points" tool which a tool that extracts the cell values of a raster based on a set of point features and records the values in the attribute table of an output feature class (Esri, n.d.). In this case the raster is the WorldlPop raster, the point features are the vector points converted from the WorldPop raster, and output feature class is the resulting vector points layer. These values are the WorldPop estimates that were used in the analysis and not the integer values. These vector points (which represent the centroids of the WorldPop 100m raster grids) were also converted to raster using "objectID" as the value field to produce a raster layer with each cell having a unique value. The "objectID" field is a field that contains the unique value that uniquely identifies a record in a layer, in this case a point and the value field is the field used in allocating values to each raster cell (Esri, n.d.). The purpose of this step was to ensure that each cell had its own value, thus retaining all the grids after converting the raster layer to vector polygons to avoid dissolving them (which is what happens when the WorldPop raster layer is converted to polygons). The raster layer was then converted to polygons, resulting in a gridded vector polygon layer. A spatial join was then performed between the resultant polygon layer and the vector points layers for pregnancies and births that had the floating values for each pregnancies and live births record. A spatial join transfers the attribites from one feature class to another based on the spatial relationships between the features in the two feature classes (Esri, n.d.). The resulting layer was a vector grids layer with floating values of WorldPop pregnancies and live births. This layer was then joined with the GPS points layer using a spatial join with the merge rule "sum" for the births and pregnancies fields to produce a layer of 100m vector polygon grids whose attribute table was then exported to Excel for analysis. The merge rule is used to specify what you want to do if there is more than one feature in the feature class that is being joined (GPS coordinates) overlaying with one feature in the target feature class (vector grids). In this case since more than one GPS points would be encountered within one grid in many instances, the sum of the values of those points for both births and pregnancies values would be returned for each grid. Figure 3.1

shows an overlay of the three feature classes that were joined to produce the final table.

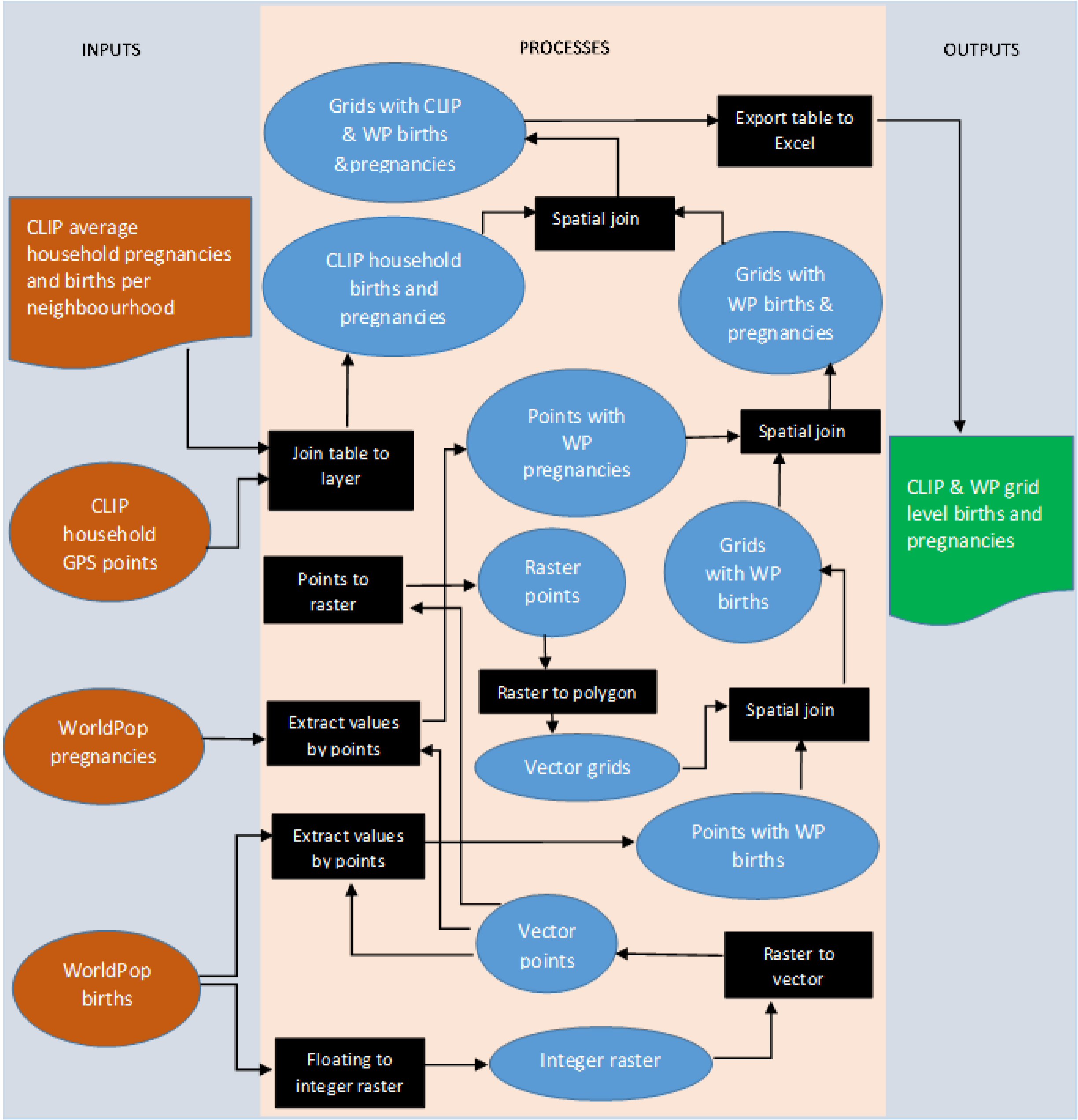

Untitled - ArcMap - ArcInfo
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:3,000
Classification ▾ Layer:
Editor ▾
Table Of Contents
Layers
C:\Users\yolis\OneDrive\project\
Coordinates
C:\Users\yolis\OneDrive\project\
rastert_pregs1_Clip
preg_sj
ArcToolbox
Catalog
Search
32.652 -25.021 Decimal Degrees

Source: Author

Figure 3.1: Overlay of the three layers, vector grids (square grids), WorldPop estimates (red points) and CLIP GPS points (blue points)

As seen in the figure above, WorldPop estimates are centroids of the grids and there are more than one features of the GPS household points occurring within some of the grids.

Figure 3.2 below shows a schematic illustration of the whole process of producing data that was used for analysis of the WorldPop maternal and perinatal datasets.

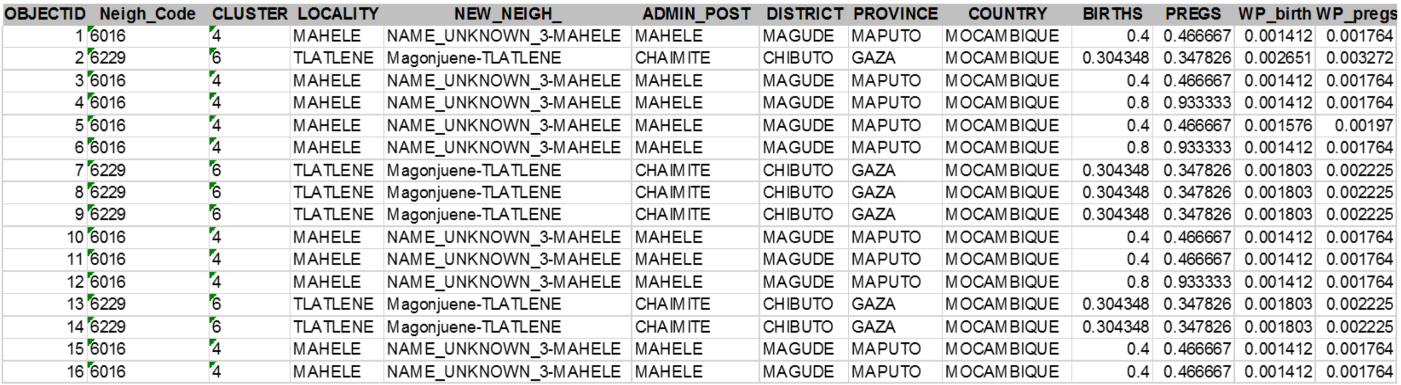


Source: Author

Figure 3.2: Flow diagram for inputs, processes and outputs used in data processing

The output table that was exported to Excel is shown as table 3.1 below showing only a few records.

Table 4.1: Resultant table showing the data used for analysis

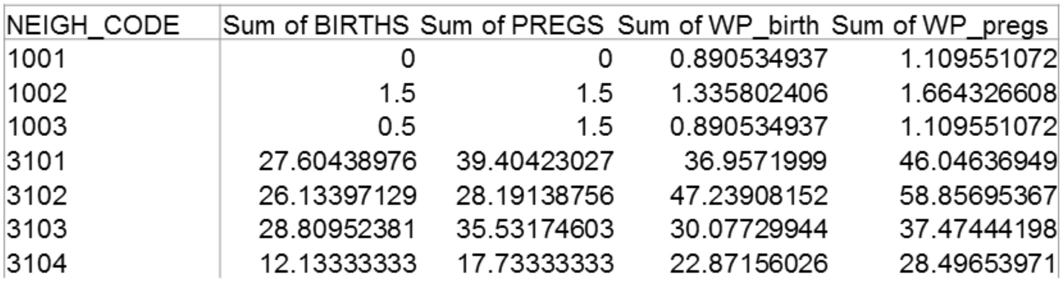


OBJECTID	Neigh_Code	CLUSTER	LOCALITY	NEW_NEIGH_	ADMIN_POST	DISTRICT	PROVINCE	COUNTRY	BIRTHS	PREGS	WP_birth	WP_pregs
1	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.4	0.466667	0.001412	0.001764
2	6229	6	TLATLENE	Magonjuene-TLATLENE	CHAIMITE	CHIBUTO	GAZA	MOCAMBIQUE	0.304348	0.347826	0.002651	0.003272
3	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.4	0.466667	0.001412	0.001764
4	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.8	0.933333	0.001412	0.001764
5	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.4	0.466667	0.001576	0.00197
6	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.8	0.933333	0.001412	0.001764
7	6229	6	TLATLENE	Magonjuene-TLATLENE	CHAIMITE	CHIBUTO	GAZA	MOCAMBIQUE	0.304348	0.347826	0.001803	0.002225
8	6229	6	TLATLENE	Magonjuene-TLATLENE	CHAIMITE	CHIBUTO	GAZA	MOCAMBIQUE	0.304348	0.347826	0.001803	0.002225
9	6229	6	TLATLENE	Magonjuene-TLATLENE	CHAIMITE	CHIBUTO	GAZA	MOCAMBIQUE	0.304348	0.347826	0.001803	0.002225
10	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.4	0.466667	0.001412	0.001764
11	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.4	0.466667	0.001412	0.001764
12	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.8	0.933333	0.001412	0.001764
13	6229	6	TLATLENE	Magonjuene-TLATLENE	CHAIMITE	CHIBUTO	GAZA	MOCAMBIQUE	0.304348	0.347826	0.001803	0.002225
14	6229	6	TLATLENE	Magonjuene-TLATLENE	CHAIMITE	CHIBUTO	GAZA	MOCAMBIQUE	0.304348	0.347826	0.001803	0.002225
15	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.4	0.466667	0.001412	0.001764
16	6016	4	MAHELE	NAME_UNKNOWN_3-MAHELE	MAHELE	MAGUDE	MAPUTO	MOCAMBIQUE	0.4	0.466667	0.001412	0.001764

3.4 ANALYSIS OF THE DATA

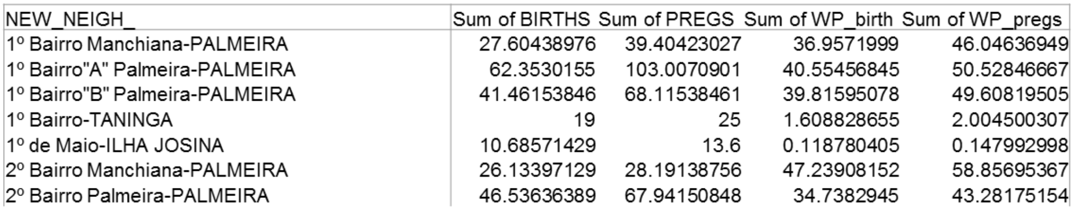
For each of the administrative unit levels created from the GPS coordinates in the study, corresponding values for each of the pregnancies and births outcomes was generated using the pivot table tool. The pivot table tool in Excel is used to arrange and summarise data. This tool was used to produce six different tables, one for each administrative level. The tables are shown below and only the first 7 entries will be shown.

Table 3.2: Tables a) to f) showing results at each administrative level from the highest level table a) neighbourhood to the lowest level table f) district



NEIGH_CODE	Sum of BIRTHS	Sum of PREGS	Sum of WP_birth	Sum of WP_pregs
1001	0	0	0.890534937	1.109551072
1002	1.5	1.5	1.335802406	1.664326608
1003	0.5	1.5	0.890534937	1.109551072
3101	27.60438976	39.40423027	36.9571999	46.04636949
3102	26.13397129	28.19138756	47.23908152	58.85695367
3103	28.80952381	35.53174603	30.07729944	37.47444198
3104	12.13333333	17.73333333	22.87156026	28.49653971

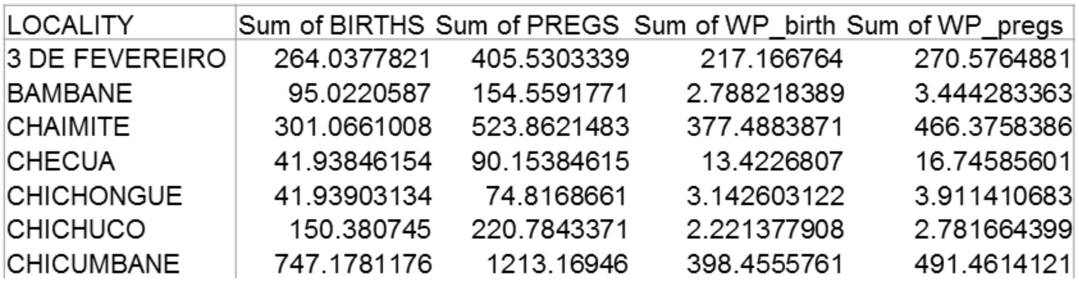
a)



NEW_NEIGH_	Sum of BIRTHS	Sum of PREGS	Sum of WP_birth	Sum of WP_pregs
1º Bairro Manchiana-PALMEIRA	27.60438976	39.40423027	36.9571999	46.04636949
1º Bairro"A" Palmeira-PALMEIRA	62.3530155	103.0070901	40.55456845	50.52846667
1º Bairro"B" Palmeira-PALMEIRA	41.46153846	68.11538461	39.81595078	49.60819505
1º Bairro-TANINGA	19	25	1.608828655	2.004500307
1º de Maio-ILHA JOSINA	10.68571429	13.6	0.118780405	0.147992998
2º Bairro Manchiana-PALMEIRA	26.13397129	28.19138756	47.23908152	58.85695367
2º Bairro Palmeira-PALMEIRA	46.53636389	67.94150848	34.7382945	43.28175154

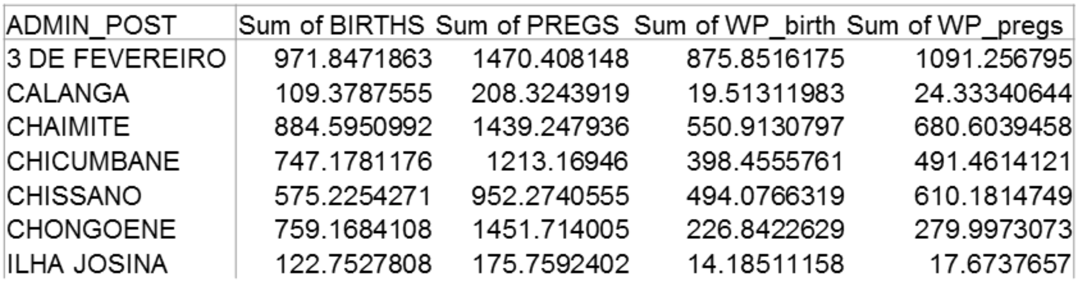
b)

c)



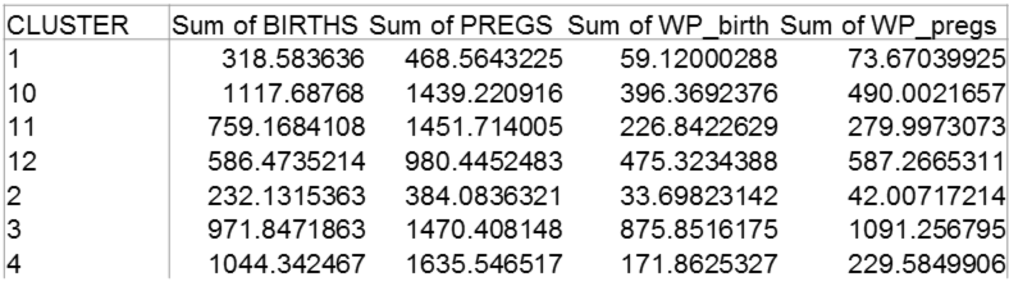
LOCALITY	Sum of BIRTHS	Sum of PREGS	Sum of WP_birth	Sum of WP_pregs
3 DE FEVEREIRO	264.0377821	405.5303339	217.166764	270.5764881
BAMBANE	95.0220587	154.5591771	2.788218389	3.444283363
CHAIMITE	301.0661008	523.8621483	377.4883871	466.3758386
CHECUA	41.93846154	90.15384615	13.4226807	16.74585601
CHICHONGUE	41.93903134	74.8168661	3.142603122	3.911410683
CHICHUCO	150.380745	220.7843371	2.221377908	2.781664399
CHICUMBANE	747.1781176	1213.16946	398.4555761	491.4614121

d)



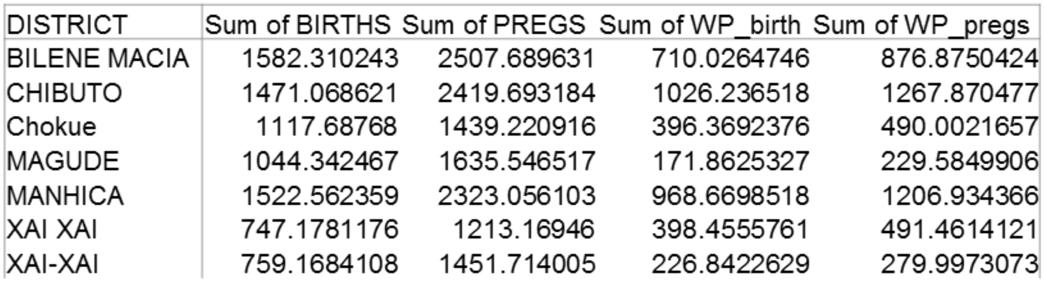
ADMIN_POST	Sum of BIRTHS	Sum of PREGS	Sum of WP_birth	Sum of WP_pregs
3 DE FEVEREIRO	971.8471863	1470.408148	875.8516175	1091.256795
CALANGA	109.3787555	208.3243919	19.51311983	24.33340644
CHAIMITE	884.5950992	1439.247936	550.9130797	680.6039458
CHICUMBANE	747.1781176	1213.16946	398.4555761	491.4614121
CHISSANO	575.2254271	952.2740555	494.0766319	610.1814749
CHONGOENE	759.1684108	1451.714005	226.8422629	279.9973073
ILHA JOSINA	122.7527808	175.7592402	14.18511158	17.6737657

e)



CLUSTER	Sum of BIRTHS	Sum of PREGS	Sum of WP_birth	Sum of WP_pregs
1	318.583636	468.5643225	59.12000288	73.67039925
10	1117.68768	1439.220916	396.3692376	490.0021657
11	759.1684108	1451.714005	226.8422629	279.9973073
12	586.4735214	980.4452483	475.3234388	587.2665311
2	232.1315363	384.0836321	33.69823142	42.00717214
3	971.8471863	1470.408148	875.8516175	1091.256795
4	1044.342467	1635.546517	171.8625327	229.5849906

f)



DISTRICT	Sum of BIRTHS	Sum of PREGS	Sum of WP_birth	Sum of WP_pregs
BILENE MACIA	1582.310243	2507.689631	710.0264746	876.8750424
CHIBUTO	1471.068621	2419.693184	1026.236518	1267.870477
Chokue	1117.68768	1439.220916	396.3692376	490.0021657
MAGUDE	1044.342467	1635.546517	171.8625327	229.5849906
MANHICA	1522.562359	2323.056103	968.6698518	1206.934366
XAI XAI	747.1781176	1213.16946	398.4555761	491.4614121
XAI-XAI	759.1684108	1451.714005	226.8422629	279.9973073

For this research two regression analysis methods were used to assess the accuracy of the WorldPop estimates, the RMSE and coefficient of variance (CV). RMSE was used to measure the errors of values for pregnancies and births predicted by WorldPop using the values observed from the CLIP baseline census.

3.4.1 ROOT MEAN SQUARE ERROR (RMSE)

The RMSE has been mostly used by different authors like (Lung et al., 2013), (Douglass et al., 2015), (Cockx and Canters, 2015) and others to examine model-performance of the population datasets they created. The root mean square error (RMSE) is a standard statistical metric used to measure model performance (Chai and Draxler, 2014) that has been applied in many disciplines. It has been shown that it is more appropriate to use RMSE over mean absolute error MAE when the model errors follow a normal distribution (Chai and Draxler, 2014). This conclusion came upon when Chai and Draxler, (2014) contended claims made by Willmott and Matsuura, (2005) that argued that RMSE does not perform well as an indicator of average model performance which were seen to be appreciated and quoted by many authors. In this research, it has been shown that the variables, pregnancies and births, follow a normal distribution at aggregated levels (shown in figure 4.2) thus indicating that RMSE was the more appropriate choice for analysing model performance. The variables do not follow a normal distribution at grid cell level therefore RMSE was not used for statistical analysis at this level.

The formula for RMSE is given below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_{obs,i} - P_{model,i})^2}{n-1}}$$

Where: P_{obs} is the number of births or pregnancies observed from the baseline data within each administrative unit.

P_{model} is the number of births or pregnancies modelled by WorldPop within same administrative unit.

n is the number of administrative units under a certain administrative unit level.

For each dataset (births and pregnancies) RMSEs were calculated at each administrative level, that is neighbourhood, new neighbourhood, locality, administrative post, cluster and district level. The coefficients of variance were also calculated using the equation below.

$$CV = RMSE^2$$

$$CV = \frac{\sum_{i=1}^{n}(P_{obs,i} - P_{model,i})^2}{n-1}$$

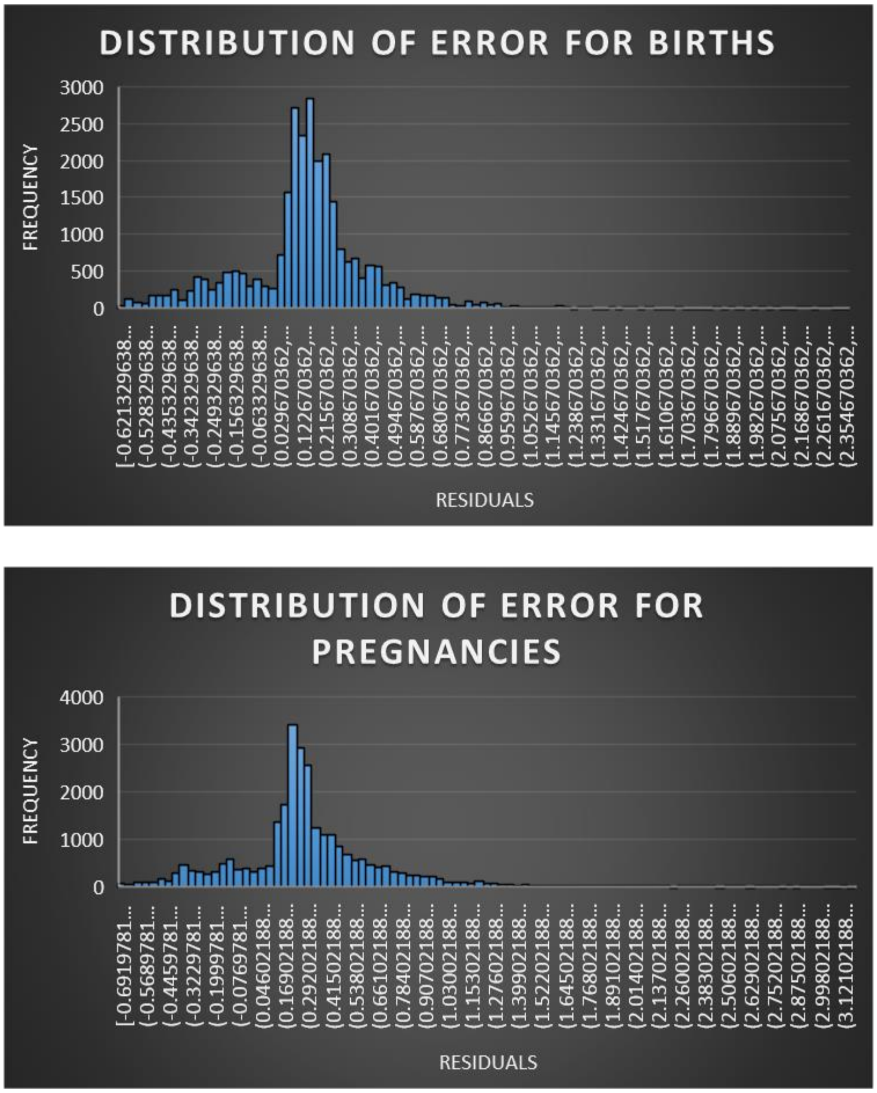
Percentage RMSE (% RMSE), also referred to as the normalised RMSE, analysis was used to allow the comparison of the performance of the model at the different administrative unit levels (Hay et al., 2005; Patel et al., 2016). % RMSE is the ratio of the RMSE to the average of the observed values expressed as a percentage. The % RMSE for pregnancies and births was computed by dividing the RMSE by the average CLIP outcomes for pregnancies and births. Percentages between 0-100% indicate an underestimation and percentages greater than 100% indicate an overestimation. Lower percentages indicate less residual variance and higher percentages indicate more variance between the modelled and the observed values within a sample.

$$\%RMSE = \frac{RMSE}{\overline{P_{obs,\iota}}}$$

Where: $\overline{P_{obs,\iota}}$ is the average count of pregnancies or live births at the i^{th} administrative unit level.

3.4.2 WELCH’S t-TEST FOR UNEQUAL VARIANCES

Despite its under-utilisation when compared to the Students t-Test and the Mann-Whitney tests the Welch’s t-Test has been proven to perform as well as or better than these two methods in terms of control of Type I (incorrect rejection of a true null hypothesis) and Type II (incorrect retaining of a false null hypothesis) (Ruxton, 2006). The t- test was performed on Excel using the “t-Test: Two-Sample Assuming Unequal Variances” method. This test was done to determine if there is sufficient evidence to conclude that there is significant difference in the performance of the WorldPop dataset in estimating births from estimating pregnancies. The test was used to test if the residual means from two samples (CLIP and WorldPop) are equal or not. The null hypothesis states that the means are equal and hence there is no significant difference in the performance of WorldPop in modelling births and in modelling pregnancies. The test was done on the residuals at grid cell level at a 95% confidence level. The rejection criterion for this test is, if the calculated P-value is less than the defined level of significance, in this case 0.05 and the t-Statistic is also less than the critical t-value then we reject the null hypothesis and conclude that there is a difference in the average errors for the birth dataset and the pregnancies dataset. The two-tail test was used as the residuals exhibited an approximate two tailed distribution as shown in figure 3.3.



Source: Author

Figure 3.3: Distributions of errors for births (top) and pregnancies (bottom) datasets.

3.4.3 PEARSON'S CORRELATION COEFFICIENT (R) AND THE COEFFICIENT OF DETERMINATION ($\boldsymbol{R^2}$)

Pearson's correlation coefficient was used to measure the degree to which the changes in the CLIP outcomes within an administrative unit were predicted by the WorldPop estimates. A positive value would mean a positive linear relationship between the observed and the modelled meaning that the variation of the observed values in a certain direction is reflected by the model. The coefficient of determination was used to determine the proportion of fluctuation of CLIP outcomes that are predictable from the WorldPop estimates. This measure allowed us to determine how certain we can be in making variation predictions using the WorldPop model. The coefficient of determination allowed the determination of the magnitude of fraction of the variances between WorldPop and CLIP values was described by the linear fit. The value falling within the range 0-1 reflects the percentage of the modelled values that can be explained by the shown linear relationship between the observed and modelled values. Of note is the fact that a linear relationship is assumed in this case. This may not be the case since there were many factors involved in modelling the distribution of these values influencing their outcomes. The coefficient of determination is calculated by squaring the correlation coefficient.

$$r = \frac{\sum_{i=1}^{n}\left(P_{obs,i} - \overline{P_{obs}}\right) * \left(P_{model,i} - \overline{P_{model}}\right)}{\sqrt{\sum_{i=1}^{n}\left(P_{obs,i} - \overline{P_{obs}}\right)^2 * \left(P_{model,i} - \overline{P_{model}}\right)^2}}$$

Where $\overline{P_{obs}}$ and $\overline{P_{model}}$ are the average values for the observed CLIP pregnancies or births and WorldPop's estimates respectfully.

3.5 DETERMINING PRECISION OF WORLDPOP DATA

This process was done to determine by how much the WorldPop model predicted populations to exist when per the CLIP census there are no populations. This was done for the district and locality levels as the data was available from the DivaGIS website. The WorldPop datasets were converted to vector points and the points were used to extract the raster values from the births and pregnancies datasets. A spatial join was done between the districts and the locality polygons to obtain the WorldPop values for pregnancies and births. The final layer produced from the grid cell analysis was converted to points and a spatial join was done between the resultant points layer and the districts and locality polygons to produce a final layer. The attributes of interest of the layer are the CLIP outcomes, WorldPop estimates where there is occurrence of household GPS points and the WorldPop estimates within the entire polygon for each administrative unit. The percentages of mismatch were calculated by dividing the difference between WorldPop's total estimates within the entire boundary and estimates at habited areas by the total estimates within the administrative unit and expressing this ration as a percentage. The percentages are translated by saying of the estimates made by the WorldPop model the percentage value obtained is the amount by which the WorldPop misidentified the populated areas. The higher the percentage the poorer the WorldPop's births and pregnancies data sets' performance in accurately assigning values to the populated areas.

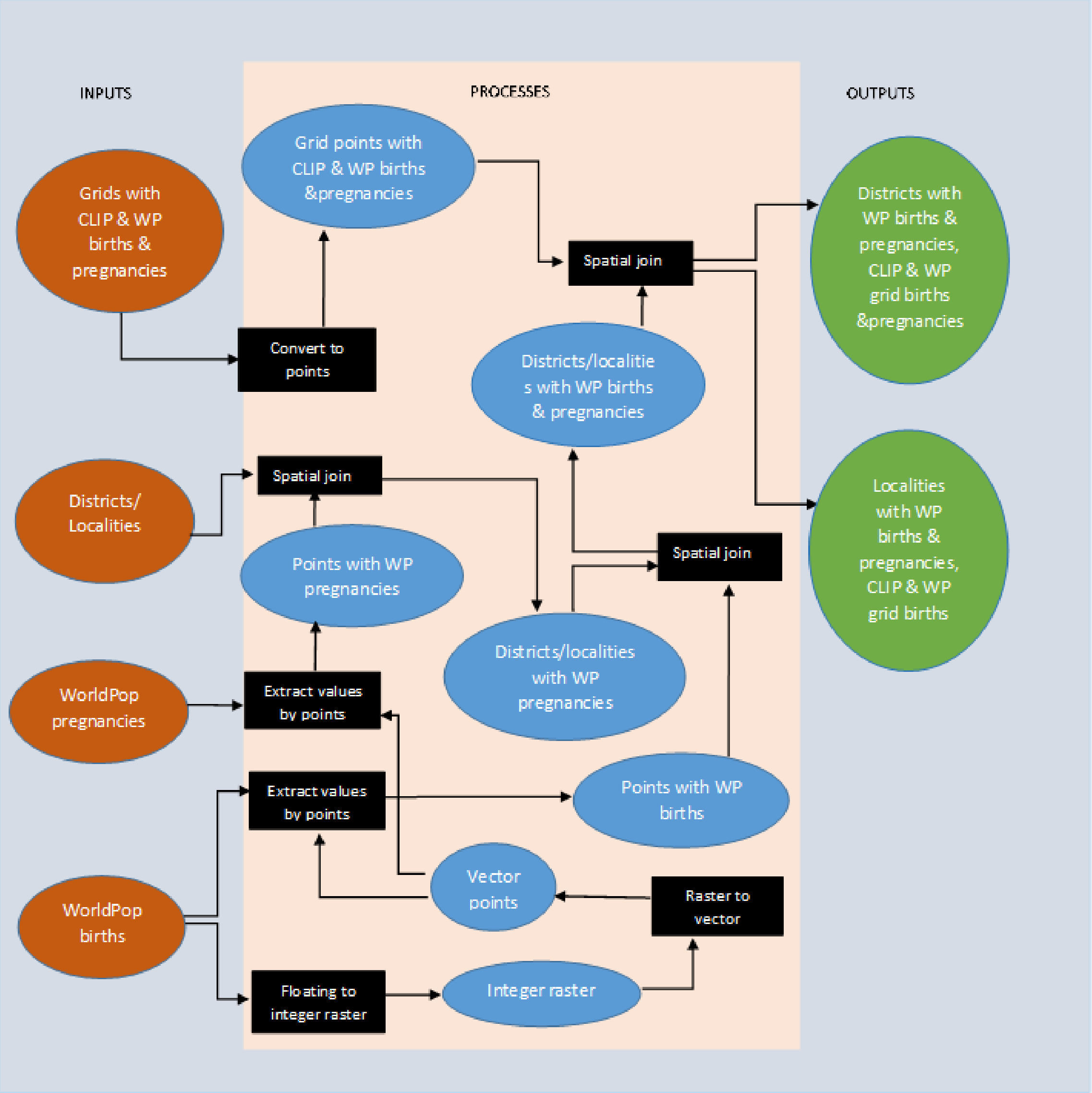

INPUTS
PROCESSES
OUTPUTS
Grids with CLIP & WP births & pregnancies
Grid points with CLIP & WP births &pregnancies
Convert to points
Spatial join
Districts/localitie s with WP births & pregnancies
Districts with WP births & pregnancies, CLIP & WP grid births &pregnancies
Districts/ Localities
Spatial join
Points with WP pregnancies
Spatial join
Localities with WP births & pregnancies, CLIP & WP grid births
Districts/localities with WP pregnancies
WorldPop pregnancies
Extract values by points
Extract values by points
Points with WP births
Vector points
Raster to vector
WorldPop births
Floating to integer raster
Integer raster

Figure 3.4: Flow diagram for inputs, processes and outputs used in data processing for determination of percentage mismatch.

CHAPTER 4: RESULTS

4.1 INTRODUCTION

This chapter presents the findings from the investigations performed in the previous chapter. Included are the results from the grid cell level analysis that include the RMSE, CV, % overestimation, % underestimation and the Welch’s unequal variance t-test results. The analyses of different levels of administrative units produced results for RMSE, CV, % overestimation, % underestimation and % RMSE for all the administrative levels involved in the analysis.

The summary table below shows the number of units within each administrative unit level and the total counts of births and pregnancies from both the observed and model values. The totals indicate that the WorldPop datasets underestimated the counts for live births and pregnancies by 52.71% and 62.72% respectively.

Table 4.1: Summary table for counts of live births and pregnancies.

Number of units within administrative unit level						**Total counts**			
						CLIP		**WorldPop**	
Neighbourhood code	**New neighbourhood**	**Locality**	**Administrative post**	**Cluster**	**District**	**Births**	**Pregnancies**	**Births**	**Pregnancies**
372	365	36	17	12	7	8245	12991	3899	4843

4.2 GRID CELL LEVEL DESCRIPTIVE ANALYSIS

The figure below (figs 4.1a and 4.1b) shows the comparison between WorldPop and CLIP values recorded in each grid cell. WorldPop cell values take mainly eight values as it can be seen on the chart below, although there are more than eight values. From the graphs in figure 4.1 there is no linear relationship between the values from the CLIP and WorldPop values, implying that the correlation coefficient and the coefficient of determination could not be used for analysis. Also revealed are the small ranges of values taken by the WorlPop model against a varied range of observed values. Table 4.2 shows the overall percentage overestimation and underestimation of the CLIP outcomes by the WorldPop model. The results show a high percentage underestimation.

Table 4.2: Grid cell level analysis results

	%OVERESTIMATION	%UNDERESTIMATION
BIRTHS	19.2278	80.7722
PREGNANCIES	15.9811	84.0189

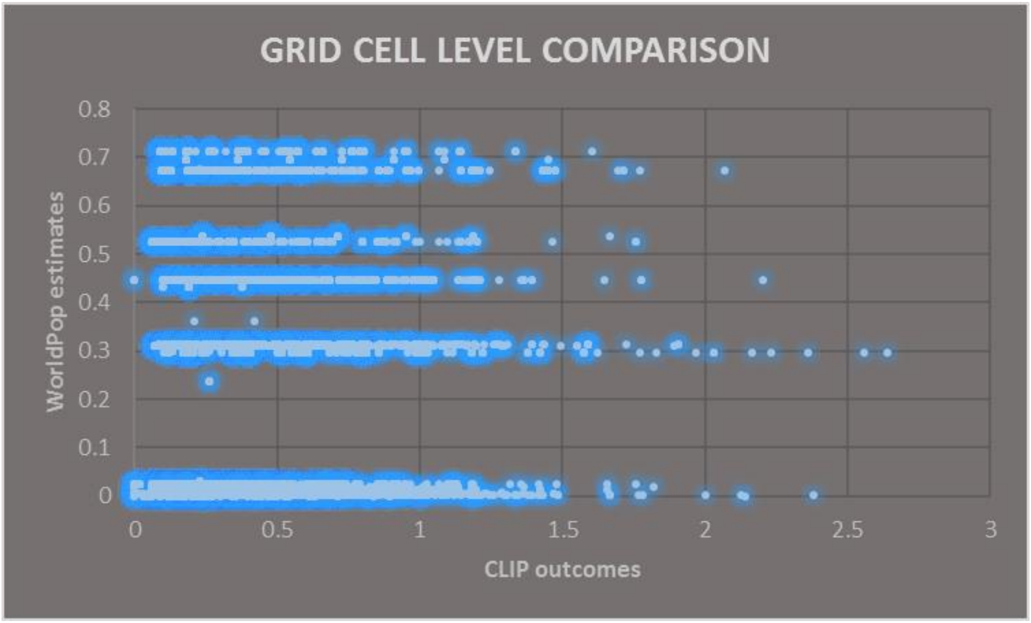


Figure 4.1a: Grid cell level distribution of WorldPop estimates and CLIP births outcomes

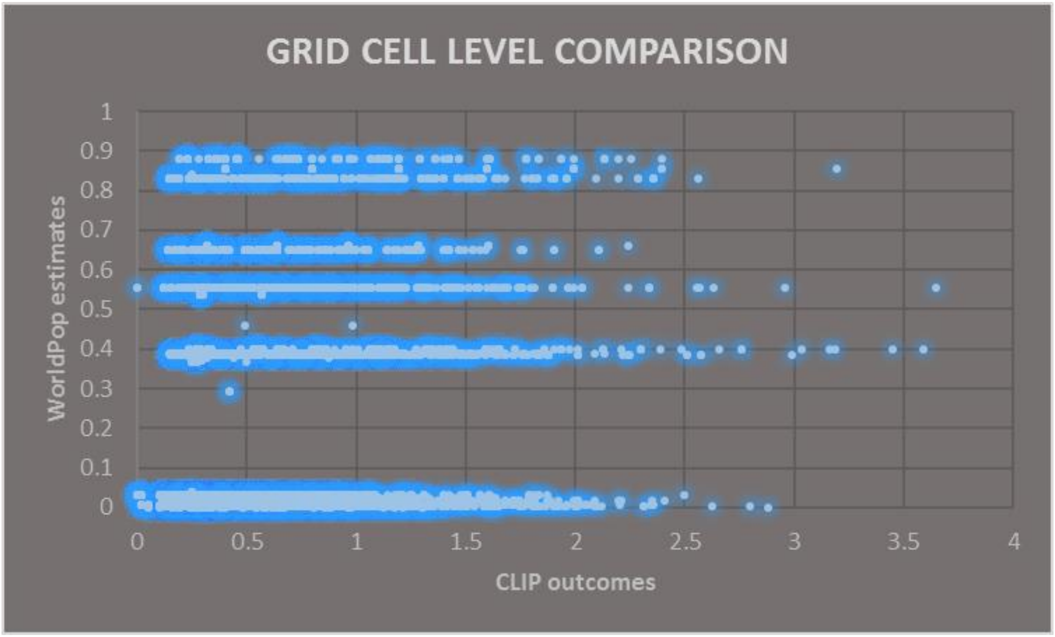


Figure 4.1b: Grid cell level distribution of WorldPop estimates and Clip pregnancies outcomes

The results of the t-test performed to determine if there is conclusive evidence that the means of the residuals of the pregnancies and live births outcomes were significantly different are shown in table 4.3.

Table 4.3: Welch's unequal variances t- Test.

t-Test: Two-Sample Assuming Unequal Variances

	birth_res	*preg_res*
Mean	0.247918711	0.374462908
Variance	0.035805645	0.074771635
Observations	27351	27351
Hypothesized Mean Difference	0	
Df	48658	
t Stat	-62.93552885	
P(T<=t) one-tail	0	
t Critical one-tail	1.644884943	
P(T<=t) two-tail	0	
t Critical two-tail	1.96001274	

As seen in the table the mean of residuals of the pregnancies is higher than the mean of the births residuals. The P-value for the two-tail test 0 is much lower than the 0.05 cut off used and the t statistic is lower than the critical t value hence the difference between the mean residuals of the two WorldPop datasets is significantly different at a 95% confidence level.

4.3 ADMINISTRATIVE UNIT LEVEL STATISTICAL ANALYSIS

The maps in figures 4.2 and 4.3 below show the distributions of the observed (CLIP) counts of live births and pregnancies at different administrative levels, illustrated by the pie charts and the bar charts within the maps. The general trend shown by the charts is that the WorldPop estimates are lower than the observed CLIP outcomes. The difference becomes more apparent with lower level administrative unit levels. At higher administrative unit levels, there are apparent variations of livebirths and pregnancies for both observed CLIP data and modelled WorldPop data. The variations become less evident in lower level administrative unit levels as the gradient of variation becomes gentler. The polygons represent the different districts covering the study area and the %underestimation at each district. The lighter regions represent a lower %underestimation range, the lowest being 40% - 50% and the darkest regions represent a 90% - 100% range. The pie charts however show that there are areas within each district where the model overestimates the observed values. The maps show a general decrease in the percentage RMSE as we move from higher to lower level administrative unit levels.

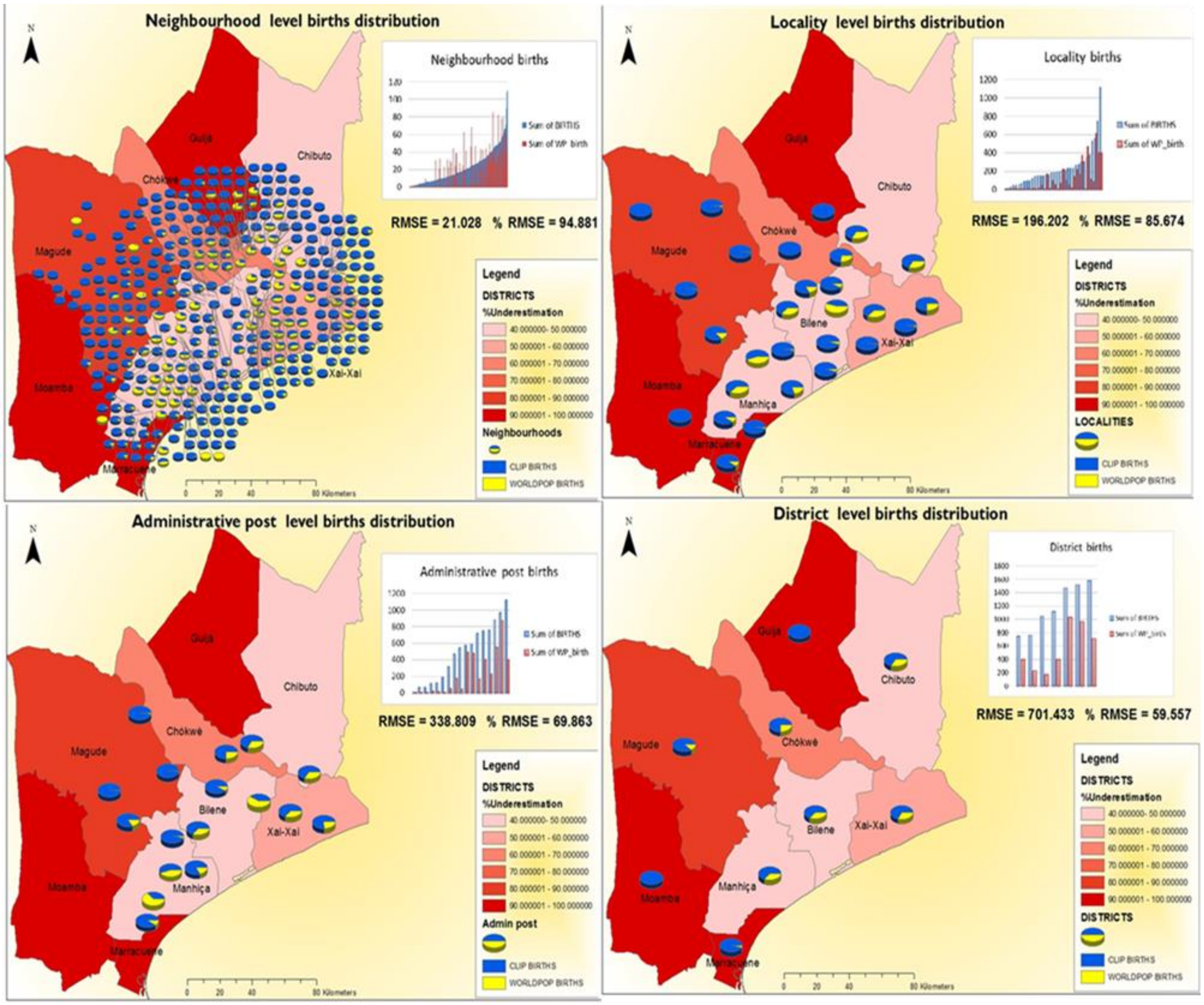


Figure 4.2: Maps showing the relationship between WorldPop estimates and CLIP outcomes of births at different administrative unit levels and %underestimation at different districts.
Source: Author

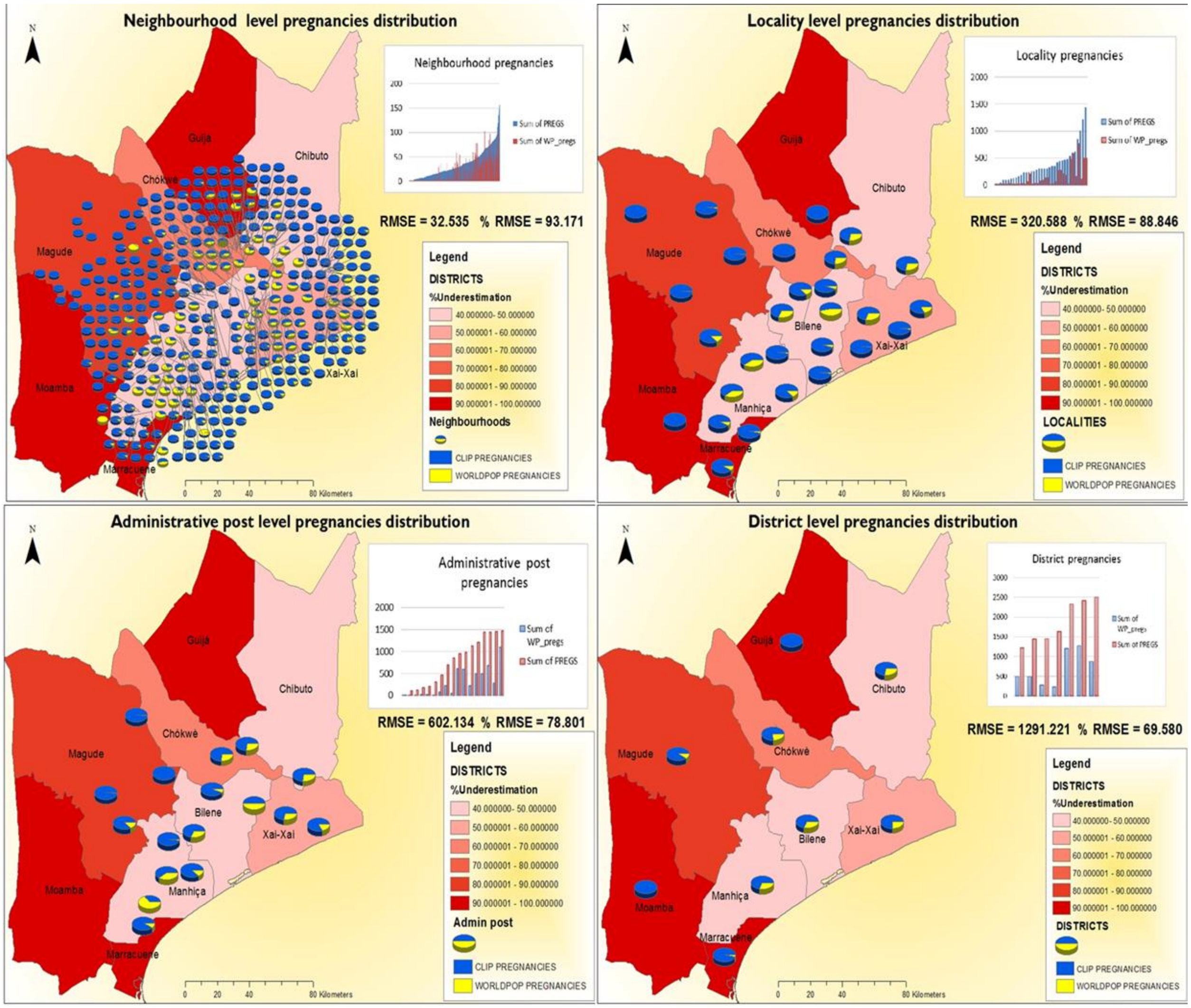


Figure 4.3: Maps showing the relationship between WorldPop estimates and CLIP outcomes of pregnancies at different administrative unit levels and %underestimation at different districts.
Source: Author

Table 4.3 shows the CV, percentage overestimation and underestimation of the WorldPop model for pregnancies and live births outcomes. The CVs increase for lower level administrative units. Percentage overestimation of the WorldPop model decreases with lower level administrative units and the percentage underestimation increases, with the single exception of locality level. At cluster and district levels the model exhibits 100% underestimation. Compared with births % RMSE for pregnancies is lower at the highest administrative unit levels and higher at lower administrative unit levels. The descriptive analysis results of the variations within different administrative units are shown in tables 4.6 – 4.9.

Table 4.3: Summary of evaluation results at different administrative unit levels

ADMINISTRATIVE LEVEL	CV		%OVER-ESTIMATION		%UNDER-ESTIMATION	
	BIRTHS	PREGS	BIRTHS	PREGS	BIRTHS	PREGS
NEIGHBOURHOOD	442.162	1058.518	13.710	7.796	86.290	92.204
NEW NEIGHBOURHOOD	458.407	1100.342	13.699	7.671	86.027	92.055
LOCALITY	38495.17	102776.9	16.667	2.778	83.333	97.222
ADMINISTRATIVE POST	114791.4	362565.1	5.882	5.882	94.118	94.118
CLUSTER	206091.1	628683.7	0	0	100	100
DISTRICT	492008.6	1667251	0	0	100	100

In tables 4.6 to 4.9 the negative percentages indicate an overestimation and positive percentages indicate an underestimation by the WorldPop model. The tables show the ranges of percentage underestimation and overestimation and number & percentage of units within each administrative unit level whose percentage overestimation or underestimation falls within each range. As seen in the tables the units within cluster and district levels do not exhibit any overestimation.

The tables show that from neighbourhood to the new neighbourhood level the highest percentage (>50%) of the units within each administrative unit underestimated by an amount between 90% and 100%. This means that at those administrative levels more than 50% of the time the model estimated 0-10% of the actual outcomes. These units are the major contributors to the overall underestimation percentage. For locality to administrative post level the percentage of units where the WorldPop model underestimated the outcomes decreases to between 20% and 30% and the percentage of areas that overestimated decreases. For cluster and district level there is no overestimation and there is an almost even distribution of percentage distributions. There is an increased frequency of units that underestimate by almost 50% of the observed outcomes.

Table 4.6: Neighbourhood level results for descriptive analysis

%DIFFERENCE	No OF NEIGHBOURHOOD UNITS		PERCENTAGE OF UNITS	
	BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES
-100 < R ≤ -90	11	4	2.957	1.075
-90 < R ≤ -80	2	0	0.538	0
-80 < R ≤ -70	2	3	0.538	0.806
-70 < R ≤ -60	1	3	0.269	0.806
-60 < R ≤ -50	3	1	0.806	0.269
-50 < R ≤ -40	2	2	0.538	0.538
-40 < R ≤ -30	4	2	1.075	0.538
-30 < R ≤ -20	3	3	0.806	0.806
-20 < R ≤ -10	2	8	0.538	2.151
-10 < R ≤ 0	9	3	2.419	0.806
0 < R ≤ 10	8	8	2.151	2.151
10 < R ≤ 20	17	7	4.57	1.882
20 < R ≤ 30	10	14	2.688	3.763
30 < R ≤ 40	12	20	3.226	5.376
40 < R ≤ 50	9	17	2.419	4.57
50 < R ≤ 60	10	11	2.688	2.957
60 < R ≤ 70	8	10	2.151	2.688
70 < R ≤ 80	5	7	1.344	1.882
80 < R ≤ 90	39	20	10.484	5.376
90 < R ≤ 100	203	229	54.57	61.559

Table 4.7: Locality level results for descriptive analysis

%DIFFERENCE	No OF LOCALITY UNITS		PERCENTAGE OF UNITS	
	BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES
-100 < R ≤ -90	0	0	0	0
-90 < R ≤ -80	0	0	0	0
-80 < R ≤ -70	0	0	0	0
-70 < R ≤ -60	0	0	0	0
-60 < R ≤ -50	1	0	2.778	0
-50 < R ≤ -40	0	0	0	0
-40 < R ≤ -30	1	0	2.778	0
-30 < R ≤ -20	1	1	2.778	2.778
-20 < R ≤ -10	0	0	0	0
-10 < R ≤ 0	3	0	8.333	0
0 < R ≤ 10	0	2	0	5.556
10 < R ≤ 20	1	2	2.778	5.556
20 < R ≤ 30	0	0	0	0
30 < R ≤ 40	2	2	5.556	5.556
40 < R ≤ 50	3	0	8.333	0
50 < R ≤ 60	0	4	0	11.111
60 < R ≤ 70	5	2	13.889	5.556
70 < R ≤ 80	2	4	5.556	11.111
80 < R ≤ 90	3	4	8.333	11.111
90 < R ≤ 100	14	15	38.889	41.667

Table 4.8: Administrative post level results for descriptive analysis

%DIFFERENCE	No OF ADMIN POST UNITS		PERCENTAGE OF UNITS	
	BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES
-100 < R ≤ -90	0	0	0	0
-90 < R ≤ -80	0	0	0	0
-80 < R ≤ -70	0	0	0	0
-70 < R ≤ -60	0	0	0	0
-60 < R ≤ -50	1	0	5.882	0
-50 < R ≤ -40	0	0	0	0
-40 < R ≤ -30	0	0	0	0
-30 < R ≤ -20	0	1	0	5.882
-20 < R ≤ -10	0	0	0	0
-10 < R ≤ 0	0	0	0	0
0 < R ≤ 10	1	0	5.882	0
10 < R ≤ 20	2	0	11.765	0
20 < R ≤ 30	0	1	0	5.882
30 < R ≤ 40	1	1	5.882	5.882
40 < R ≤ 50	1	1	5.882	5.882
50 < R ≤ 60	0	2	0	11.765
60 < R ≤ 70	2	2	11.765	11.765
70 < R ≤ 80	2	0	11.765	0
80 < R ≤ 90	3	5	17.647	29.412
90 < R ≤ 100	4	4	23.529	23.529

Table 4.9: District level results for descriptive analysis

%DIFFERENCE	No OF DISTRICT UNITS		PERCENTAGE OF UNITS	
	BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES
0 < R ≤ 10	0	0	0	0
10 < R ≤ 20	0	0	0	0
20 < R ≤ 30	0	0	0	0
30 < R ≤ 40	2	0	28.571	0
40 < R ≤ 50	1	2	14.286	28.571
50 < R ≤ 60	1	1	14.286	14.286
60 < R ≤ 70	1	2	14.286	28.571
70 < R ≤ 80	1	0	14.286	0
80 < R ≤ 90	1	2	14.286	28.571
90 < R ≤ 100	0	0	0	0

4.3.1 RESULTS FROM CC AND CD ANALYSIS

Table 4.5: Linear relationship analysis results

ADMINISTRATIVE LEVEL	CORRELATION COEFFICIENT (R)		COEFFICIENT OF DETERMINATION (R^2)	
	BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES
NEIGHBOURHOOD	0.541394889	0.576658841	0.293108425	0.332535419
LOCALITY	0.715863322	0.731758555	0.512460296	0.535470583
ADMINISTRATIVE POST	0.799327747	0.800355995	0.638924848	0.640569718
DISTRICT	0.827613364	0.840672908	0.68494388	0.706730939

Table 4.5 shows the results from the analyses of the linear relationships of the WorldPop model and the observed CLIP values at different administrative levels. The results show a moderate positive relationship, from the CC values above 50%, for neighbourhood, new neighbourhood and cluster levels and a fairly-strong relationship, from CC values, above 70% for locality, administrative post and district levels. The neighbourhood, new neighbourhood and cluster levels also exhibit coefficient of determination values below 50% while the locality, administrative post and district levels have R^2 values above 50%. These values however are affected by the size of the sample as shown in figures 4.4 and 4.5. The figures below show the line graphs for the linear relationship between the model and observed values at different administrative unit levels, with the equations of the line of best fit and R^2 values. As it can be seen from the figures the sizes of the samples influence the R^2 values. At neighbourhood level, more points lie close to the line of best fit than at district level but because the ratio the number of those points to the entire sample size is higher than that of the district level, the R^2 value for the neighbourhood level becomes lower than that for district level.

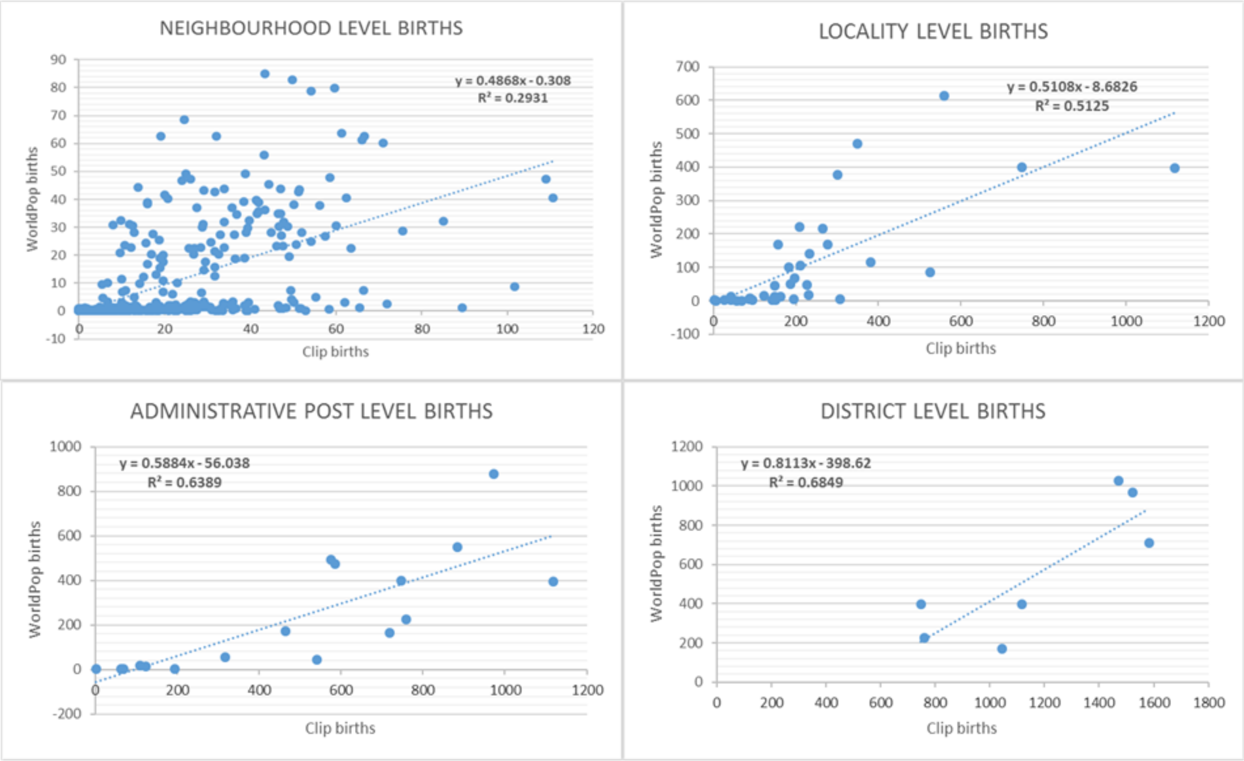


Figure 4.4: Linear relationship between WorldPop's estimates and CLIP's outcomes of births.

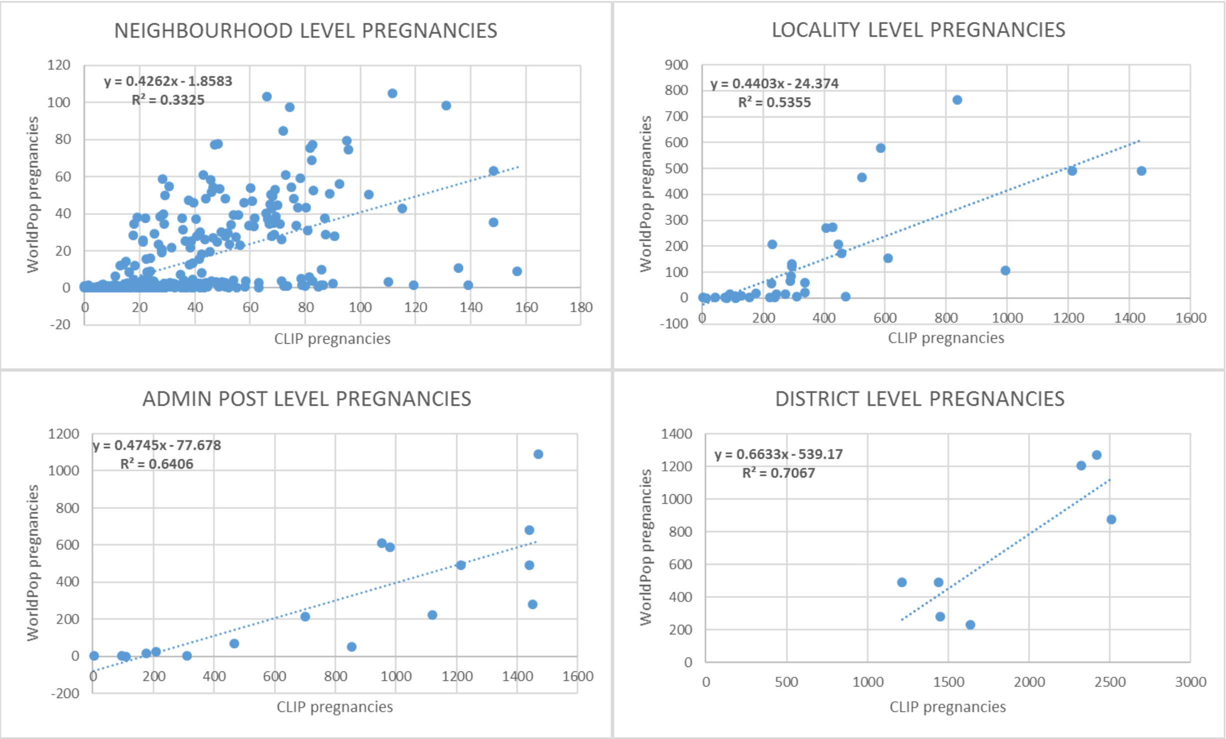


Figure 4.5: Linear relationship between WorldPop's estimates and CLIP's outcomes of pregnancies.

4.4 PERCENTAGE MISALLOCATION RESULTS

Table 4.6: Misallocation results for WorldPop live births and pregnancies datasets at district level.

BOUNDARY UNIT TOTAL		POPULATED		NON-POPULATED		%MISALLOCATION	
BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES
5600.31	6945.19	673.26	831.48	4927.04	6113.72	87.98	88.03
6850.95	8543.67	597.58	738.42	6253.38	7805.25	91.28	91.36
6691.80	8379.72	297.71	368.04	6394.09	8011.68	95.55	95.61
3134.83	3868.12	0.43	0.53	3134.40	3867.59	99.99	99.99
4500.18	5565.50	565.86	698.11	3934.32	4867.39	87.43	87.46
1747.36	2208.67	109.14	145.74	1638.22	2062.93	93.75	93.40
7181.94	8977.09	769.50	958.79	6412.44	8018.30	89.29	89.32
4140.72	5162.58	0.28	0.35	4140.44	5162.23	99.99	99.99
1950.90	2486.93	0.03	0.03	1950.87	2486.90	99.99	99.99

Table 4.7: Misallocation results for WorldPop live births and pregnancies datasets at locality level

BOUNDARY UNIT TOTAL		POPULATED		NON-POPULATED		%MISALLOCATION	
BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES	BIRTHS	PREGNANCIES
175.2	218.98	1.08	1.34	174.13	217.63	99.39	99.39
2714.15	3386.14	708.65	882.93	2005.5	2503.21	73.89	73.93
896.44	1117.32	24.73	30.85	871.71	1086.47	97.24	97.24
233.6	291.05	4.18	5.21	229.41	285.84	98.21	98.21
1206.88	1504.12	28.64	35.68	1178.24	1468.44	97.63	97.63
1237.85	1542.4	3.3	4.11	1234.54	1538.28	99.73	99.73
728.51	913.88	1.44	1.78	727.06	912.1	99.8	99.8
1599.65	1975.63	453.05	559.51	1146.61	1416.11	71.68	71.68
749.09	925.08	0.91	1.13	748.18	923.95	99.88	99.88
788.54	973.93	38.1	47.06	750.44	926.87	95.17	95.17
849.59	1049.25	179.73	221.96	669.86	827.29	78.85	78.85
884.95	1107.45	0.03	0.03	884.92	1107.42	99.99	99.99
1132	1398.39	402.25	497.03	729.75	901.35	64.47	64.46
905.33	1118.52	195.33	241.39	710	877.13	78.42	78.42
1746.56	2193.27	0.4	0.5	1746.16	2192.78	99.98	99.98
773.84	956.47	297.31	367.54	476.53	588.92	61.58	61.57
784.77	968.37	0.43	0.53	784.34	967.84	99.95	99.95
1890.05	2332.65	378.3	466.69	1511.74	1865.95	79.98	79.99
1619.76	1998.69	186.9	230.58	1432.85	1768.11	88.46	88.46
378.46	479.01	0.58	0.74	377.87	478.27	99.85	99.85
611.92	755.15	0.07	0.09	611.85	755.06	99.99	99.99
657.05	845.93	105.87	141.65	551.18	704.29	83.89	83.26
180.22	225.24	0.61	0.76	179.61	224.48	99.66	99.66
536.9	671.06	0.03	0.03	536.87	671.02	99.99	99.99
197.98	247.46	1.56	1.95	196.42	245.51	99.21	99.21
1097.72	1369.69	0.22	0.27	1097.51	1369.42	99.98	99.98
3043	3792.89	0.06	0.08	3042.93	3792.81	99.99	99.99
802.21	1034.25	0.03	0.03	802.19	1034.22	99.99	99.99

The tables 4.6 and 4.7 above show high percentages, between 80% and 100% for district level and between 70% and 100% for locality level, of the total estimates within each administrative unit being allocated in non-populated areas per the CLIP household GPS points. This means between 70% and 100% of the allocated estimates within the administrative units are allocated in areas that are not populated.

CHAPTER 5: DISCUSSION

5.1 PERFORMANCE OF WORLDPOP DATASET AT DIFFERENT SPATIAL SCALES

The decreasing amount of fluctuations of the differences between the model estimates and observed outcomes and the 100% underestimation observed at higher level administrative units are an indicator of how the model masks heterogeneity. With higher level, administrative units, less detail is revealed thus heterogeneity that exists within the units is not revealed. The increasing RMSEs might have been a consequence of errors incorporated when mapping at a large scale. The errors inherent in the methods and data used are amplified in smaller areas. Land cover data is less efficient when the land cover classes disaggregate over larger areas than the habited area whose population is to be disaggregated, thus the smaller scale information that will be contained within the sub-partitions is lost (Dmowska and Stepinski, 2014).

Accumulation of the smaller areas into larger areas therefore also means accumulation of such errors, thus as higher level administrative units are accumulated to lower level administrative units the RMSE increases. The great difference between percentage underestimation and overestimation attests the strength of the covariates and statistical model used, which determine the weights assigned to each cell (Lung et al., 2013). Resolution of the remotely sensed land cover data is one of the factors that could have played a role in the high percentage of underestimation. This is because land cover data is less efficient when the land cover classes disaggregate over larger areas than the habited are whose population is to be disaggregated, thus the smaller scale information that will be contained within the sub-partitions is lost (Dmowska and Stepinski, 2014). This is an indicator that the WorldPop model may not capture the heterogeneities within densely populated areas as some land cover classes can mask them, especially when using a medium resolution dataset.

Also, attesting the reliability of the land cover data used are the ranges of values observed at grid cell evaluation which are too few. Taking into consideration the fact that the CLIP outcomes correspond to households in different geographic locations and with different socioeconomic characteristics, the grid cell values are depicting a rather amassed representation of the distributions. Although the land cover class may be the same (residential houses) the different socioeconomic characteristics and classes of urban densities should diversify the distributions hence the grid cell values observed in the WorldPop dataset. Mennis, (2003) identified three urban classes high density, low density and non-urban classes (Jia et al., 2014). Demographic variations across these classes need to be reflected to enable detection of the vulnerable groups and individuals at risk (Tatem et al., 2014). Jia and Gaughan, (2016) suggested the use of property types derived from parcel data in combination with land cover data in assigning weights to the grid cells to increase the accuracy of datasets, instead of only using settlement data. This method also improves detail, which is shown to be low in this study by the small range of the grid values and allocation accuracy is improved as property types are better indicators of population density than land cover data (Jia et al., 2014).

5.2 MAPPING PRECISION OF THE WORLDPOP DATASET

Although quantitatively the WorldPop model showed low precision from the results of percentage misallocation, the general variation trend conforms very well to that of the observed outcomes as shown on the charts and as mentioned before. This relationship means that the WorldPop datasets can be used in identifying the most affected regions with respect to the other regions. The model conforms to the trends depicted by the observed values and the improvement of the conformity at higher level administrative boundaries indicates that at such administrative levels, the WorldPop datasets perform better in showing the relative outcomes, from unit to unit, at the same administrative unit level. This makes it possible to use the datasets in determining which regions need interventions more than the others.

Percentage RMSEs indicate that the WorldPop model performs better at higher level administrative units as they are lower at those levels. The higher %RMSE indicates that the ratio of error to observed population of live births and pregnancies within a boundary unit is very high which indicates poor performance of the model. This means that the WorldPop datasets better model the maternal and perinatal outcomes at higher administrative unit levels.

The results have shown that overall WorldPop’s live births dataset estimates the live births outcomes better than the pregnancies dataset estimates the pregnancies outcomes. These results reflect the accumulative errors inherited in the pregnancies dataset due to production of the dataset using adjusted births dataset (Tatem et al., 2014). The creation of these datasets is in three stages which have inherent errors, with the births dataset depending on the population distribution to allocate estimates in habited areas and the pregnancies dataset relying on the births dataset and national level estimates which assume nonexistence of spatial variation within the country (Tatem et al., 2014).

5.3 IMPLICATIONS OF EXISTING MAPPING METHODS

While the use of different sources of ancillary data produces datasets with different accuracies, the weighting methods have proven to influence the results of dasymetric modelling approaches. The absence of a standardized weighting approach, leading to the use of different approaches based on "heuristic rules and assumptions" (Lung et al., 2013), presents a limitation in the determination of the ancillary data source that truly produces better population distribution datasets. No studies have investigated the effect of weighting approaches on the resulting datasets, instead they have only investigated the effect of choice of ancillary data and modelling techniques. These weighting approaches need to be explored to come up with an effective standardised approach. LULC data is currently not only the most used in large scale mapping but the most convenient source of ancillary data (Lung et al., 2013; Jia et al., 2014). It remains the best ancillary data for large scale mapping at affordable costs as it is currently the only data available with a large-scale coverage. There is need to further explore the statistical methods used for weighting and the relationship between land cover and population densities. Investigations of covariates that can further explain the correlation between population density and land cover/use are a necessity. These will enable the refinement of land cover/use classes, further strengthening the relationship between population density and land cover/use.

The application of telecommunications data will play a crucial role in improving the accuracy of population distribution datasets especially when considering the long inter-census periods in low income countries. Further exploration of this method's role in addressing the limitation of extremely low temporal resolutions of the census data or the lack of census data in low income regions might lead to a solution to the problem of lack of census data (Douglass et al., 2015). The use of telecommunications data has the potential to reduce errors introduced by the extrapolation of parameterised models developed using data from other countries onto countries without data. The use of telecommunications data has the potential to reduce errors introduced by the extrapolation of parameterised models developed using data from other countries onto countries without data. The ability of WorldPop's population distribution modelling approach to allow easy incorporation of new data (Deleu et al., 2015) will allow use of telecommunications data to detect population in less densely inhabited areas (Douglass et al., 2015).

All the studies have demonstrated the desire to create datasets independent of boundary data as boundary data requires good documentation and accuracy to produce quality datasets (Patterson et al., 2007). Lack of such data especially in the developing countries presents problems in mapping hence eagerness of the authors to explore more and more methods that do not require boundary data.

CHAPTER 6: CONCLUSION, FINDINGS AND RECCOMMENDATIONS

6.1 CONCLUSION

This paper has demonstrated the performance of the WorldPop's live births and pregnancies datasets at different spatial scales. It has been shown that the performance of the model is better at lower level administrative unit levels but it masks the accumulative errors at those levels. This leads to the notion that assessment of the errors in a large-scale dataset leads to a clearer conclusion if it is done at the highest level of detail which reflects the true performance of the dataset. It is to be taken into consideration that these datasets are not only used for analyses at large spatial scales but also at finer scales of spatial resolution. As shown by the results, at higher administrative unit levels the datasets perform better but have greater inherent errors masked within them. Studies that focus on assessing accuracies of these large-scale datasets have focused on validation at higher administrative levels like districts and provinces because during those periods the application of these datasets was at large spatial scales. Currently more emphasis is being made on identifying heterogeneity within areas of concern and use of high quality data for better programmed planning, calling for magnified evaluation of the global datasets.

The global data sets' potential of producing high quality data is great. Different studies have shown that more and more methods are being unveiled, with the advent of technologies that allow location of populations in real time, that will improve these datasets providing free access high quality demographic distribution data. Availability of such data on demand will enormously improve performance of intervention programmes by reducing the amount of resources used in accumulating data from different sources to perform analyses.

The intervals between formulation of policies and implementation of interventions has a great impact on the performance outcome of such programmes. Longer intervals which might be caused by the longer time taken in modelling and analysing the performance of maternal and perinatal health care systems may lead to "stagnant" improvements. This means the rate of improvement of the health care systems will be at par with the rate of increase of the populations at risk. This is particularly evident in low income regions where population growth rates are higher (Tatem and Linard, 2011). Use of well documented, high quality global datasets is the starting point in reducing and eventually removing these "stagnant" improvements.

6.2 LIMITATIONS

This study assumed that the CLIP census household GPS points represented all the habited residential houses. No check was done to determine if all the households within the study area were mapped.

It is of great importance to note that the findings of this research are restricted to the study area of Gaza and Maputo provinces. The study area in this research was very small in comparison with the whole WorldPop dataset. This did not allow the author to make conclusive findings about the dataset but to restrict the findings to the area of study only. Extrapolation of the findings to the whole country to enable a broader analysis would imply an assumption that pregnancies and live births are only affected by spatial influence. However, for a more effective extrapolation there is need to take into consideration the socio-cultural, religious, political and socioeconomic factors. Along with the poor socio-economic status, the socio-cultural behavior of most African nations plays a dominant role in influencing their reproductive behavior thus the pregnancy outcomes (Ajiboye and Adebayo, 2012).

This is because unlike population density distribution which is influenced by spatial factors that are visible, like land cover, urban and rural developments, etc., pregnancies and live births outcomes cannot simply be derived from these factors. This limitation therefore restricted the author to evaluating the WorldPop dataset for the study area.

The lack of land cover data in this research restricted the analysis of the performance of the dasymetric method in disaggregating data within very small spatial divisions, in this case the neighbourhoods, against larger spatial divisions, in this case districts, to investigate the findings by Dmowska and Stepinski, (2014) that land cover data blocks the information within sub-divisions because of covering a larger scale against the area to be assigned values. This would enable the determination of the extent to which the WorldPop masks that information due to the use of land cover data.

In this study, the analysis was only done on grid cells that had occurrences of household GPS points, whilst administrative unit boundaries are used for policy making and planning interventions. The results obtained were from an evaluation of only the habited areas within a boundary unit and not the entire boundary unit. This method was used due to lack of access to neighbourhood unit boundaries. Of note is that the results are not reflective on the performance of the WorldPop dataset within the entire boundary units but its performance within areas that are habited as per the CLIP trial census. There are areas within the administrative units which the WorldPop datasets modelled as being populated but are unpopulated as per the CLIP trial census. This is an indication that the observed underestimations at different administrative levels are indicative of the habited areas within the unit and not the entire administrative unit. The results of the scoping review within this study are not reflective on the holistic review of existing literature on the methods of population distribution mapping. Reasons for this limitation are the lack of access to electronic databases that were intended to be used for this study and the limited time to manually search for literature in other search engines besides Google Scholar. This means that the results from the scoping review are not reflective of the entire existing literature but on the identified literature during the period of the research.

6.3 RECOMMENDATIONS

For future studies pertaining to the modelling of maternal and perinatal outcomes factors mentioned before, socio-cultural, religious, political and socioeconomic need to be taken into consideration to elucidate the findings. This is because changes in population distributions are not the only influencing factors, meaning there is need to further explore the reasons for the differences between model estimates and observed outcomes within a given period.

The development of a dataset using areal interpolation would further serve as a baseline for investigating the magnitude of influence the dasymetric method has on the model estimates. Evaluating the WorldPop dataset against a dataset disaggregated using areal interpolation would allow the determination of performance of dasymetric methods at lower administrative unit levels. Other authors like Deleu et al., (2015) have proven that when dealing with very small spatial divisions areal interpolation disaggregates census data better than dasymetric methods giving better estimates. These findings need to be investigated for the maternal and perinatal datasets.

Use of administrative unit boundaries produces more realistic results when considering the purpose of the validation. This study focused on how well the dataset identifies the populated areas and the corresponding estimates for pregnancies and births. If the purpose of the validation is to determine if the dataset is suitable for accurately identifying the regions mostly affected for purposes of planning interventions or policy making, then the use of boundary data is more appropriate. This is because policy makers are not interested in which areas are habited but in the counts within an entire boundary unit. If the focus of intervention programmes shifts to identifying the location and distribution of affected populations within the units then the outcomes of this study can be applied.

For future studies, it is therefore important to consider the purpose of validating the dataset to determine the best method of validation. Validating the dataset to determine its performance in estimating populations within an administrative unit is different from validating its performance in locating populations which is also different from validating its performance to determine adjustment errors. The last-mentioned purpose of validation is of great importance although studies have not focused a lot on it. This is because its results are applicable in using the datasets. Since a dataset may not accurately estimate the observed outcomes, the adjustment values become very important in the use of the dataset.

Documentation of such adjustment values in countries that have the data for validation and information about the regions within different countries that have the same spatial and demographic characteristics will play a big role in improving results from the dataset. This means that although the other countries may lack the data to compute adjustment values, they can use the ones computed from countries listed as having the same spatial and demographic characteristics.

REFERENCES

Ajiboye, O.E., Adebayo, K.A., 2012. Socio-cultural Factors Affecting Pregnancy Outcome Among the Ogu Speaking People of Badagry Area of Lagos State, Nigeria. Int. J. Humanit. Soc. Sci., The Special Issue on Contemporary Issues in Social Science 2, 133–144.

Alegana, V.A., Atkinson, P.M., Pezzulo, C., Sorichetta, A., Weiss, D., Bird, T., Erbach-Schoenberg, E., Tatem, A.J., 2015. Fine resolution mapping of population age-structures for health and development applications. J. R. Soc. Interface 12. doi:10.1098/rsif.2015.0073

Alex Perkins, T., Siraj, A.S., Ruktanonchai, C.W., Kraemer, M.U.G., Tatem, A.J., 2016. Model-based projections of Zika virus infections in childbearing women in the Americas. Nat. Microbiol. 1, 16126. doi:10.1038/nmicrobiol.2016.126

Arksey, H., O'Malley, L., 2005. Scoping studies: towards a methodological framework. Int. J. Soc. Res. Methodol. 8, 19–32. doi:10.1080/1364557032000119616

Armstrong, R., Hall, B.J., Doyle, J., Waters, E., 2011. "Scoping the scope" of a cochrane review. J. Public Health 33, 147–150. doi:10.1093/pubmed/fdr015

Bhaduri, B., Bright, E., Coleman, P., Urban, M.L., 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal 69, 103–117. doi:10.1007/s10708-007-9105-9

CDC, 2016. Creating an Interactive R Shiny Data Visualization Application to Facilitate Global Maternal Death Surveillance and Response (MDSR) System Reporting | CHIIC | OPHSS | CDC [WWW Document]. URL http://www.cdc.gov/ophss/chiic/projects/2016/ideas/creating-an-interactive-r-shiny-data-visualization-application.html (accessed 10.8.16).

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? Geosci. Model Dev. Discuss. 7, 1525–1534. doi:10.5194/gmdd-7-1525-2014

City Population, 2016. Mozambique: Provinces, Cities, Urban Localities & Agglomeration - Population Statistics in Maps and Charts [WWW Document]. URL http://www.citypopulation.de/Mocambique.html (accessed 10.11.16).

Cockx, K., Canters, F., 2015. Incorporating spatial non-stationarity to improve dasymetric mapping of population. Appl. Geogr. 63, 220–230. doi: 10.1016/j.apgeog.2015.07.002

Deleu, J., Franke, J., Gebreslasie, M., Linard, C., 2015. Improving AfriPop dataset with settlement extents extracted from RapidEye for the border region comprising South-Africa, Swaziland and Mozambique. Geospatial Health 10. doi:10.4081/gh.2015.336

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J., 2014. Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. 111, 15888–15893. doi:10.1073/pnas.1408439111

Dmowska, A., Stepinski, T.F., 2014. High resolution dasymetric model of U.S demographics with application to spatial distribution of racial diversity. Appl. Geogr. 53, 417–426. doi: 10.1016/j.apgeog.2014.07.003

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., Worley, B.A., 2000. LandScan: a global population database for estimating populations at risk. Photogram Eng Rem Sens 66.

Douglass, R.W., Meyer, D.A., Ram, M., Rideout, D., Song, D., 2015. High resolution population estimates from telecommunications data. EPJ Data Sci. 4. doi:10.1140/epjds/s13688-015-0040-6

Ebener, S., Guerra-Arias, M., Campbell, J., Tatem, A.J., Moran, A.C., Amoako Johnson, F., Fogstad, H., Stenberg, K., Neal, S., Bailey, P., Porter, R., Matthews, Z., 2015. The geography of maternal and newborn health: the state of the art. Int. J. Health Geogr. 14. doi:10.1186/s12942-015-0012-x

Esri, n.d. ArcGIS Desktop Help 9.3 - welcome [WWW Document]. URL http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=welcome (accessed 10.12.16).

Gallego, F.J., 2010. A population density grid of the European Union. Popul. Environ. 31, 460–473. doi:10.1007/s11111-010-0108-y

Gaughan, A.E., Stevens, F.R., Linard, C., Jia, P., Tatem, A.J., 2013. High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. PLOS ONE 8, e55882. doi: 10.1371/journal.pone.0055882

GeoHive, 2016. GeoHive - Mozambique population statistics [WWW Document]. URL http://www.geohive.com/cntry/mozambique.aspx (accessed 10.11.16).

Gething, P.W., Kirui, V.C., Alegana, V.A., Okiro, E.A., Noor, A.M., Snow, R.W., 2010. Estimating the number of paediatric fevers associated with malaria infection presenting to Africa’s public health sector in 2007. PLoS Med 7. doi: 10.1371/journal.pmed.1000301

Hay, S.I., Guerra, C.A., Gething, P.W., Patil, A.P., Tatem, A.J., Noor, A.M., Kabaria, C.W., Manh, B.H., Elyazar, I.R.F., Brooker, S.J., 2009. World malaria map: Plasmodium falciparum endemicity in 2007. PLoS Med 6. doi: 10.1371/journal.pmed.1000048

Hay, S.I., Noor, A.M., Nelson, A., Tatem, A.J., 2005. The accuracy of human population maps for public health application. Trop. Med. Int. Health 10, 1073–1086. doi:10.1111/j.1365-3156.2005. 01487.x

Jia, P., Gaughan, A.E., 2016. Dasymetric modeling: A hybrid approach using land cover and tax parcel data for mapping population in Alachua County, Florida. Appl. Geogr. 66, 100–108. doi: 10.1016/j.apgeog.2015.11.006

Jia, P., Qiu, Y., Gaughan, A.E., 2014. A fine-scale spatial population distribution on the High-resolution Gridded Population Surface and application in Alachua County, Florida. Appl. Geogr. 50, 99–107. doi: 10.1016/j.apgeog.2014.02.009

Linard, C., Gilbert, M., Snow, R.W., Noor, A.M., Tatem, A.J., 2012. Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS One 7. doi: 10.1371/journal.pone.0031743

Linard, C., Gilbert, M., Tatem, A.J., 2010. Assessing the use of global land cover data for guiding large area population distribution modelling. GeoJournal 76, 525–538. doi:10.1007/s10708-010-9364-8

Linard, C., Tatem, A.J., 2012. Large-scale spatial population databases in infectious disease research. Int. J. Health Geogr. 11, 7. doi:10.1186/1476-072X-11-7

Lung, T., Lübker, T., Ngochoch, J.K., Schaab, G., 2013. Human population distribution modelling at regional level using very high resolution satellite imagery. Appl. Geogr. 41, 36–45. doi: 10.1016/j.apgeog.2013.03.002

Mennis, J., 2003. Generating Surface Models of Population Using Dasymetric Mapping. Prof. Geogr. 55, 31–42. doi:10.1111/0033-0124.10042

Neal, S., Ruktanonchai, C., Chandra-Mouli, V., Matthews, Z., Tatem, A.J., 2016. Mapping adolescent first births within three east African countries using data from Demographic and Health Surveys: exploring geospatial methods to inform policy. Reprod. Health 13. doi:10.1186/s12978-016-0205-1

Nieves, J., 2016. Global population distributions and the environment: discerning observed global and regional patterns. Electron. Theses Diss. doi: http://dx.doi.org/10.18297/etd/2427

Patel, N.N., Angiuli, E., Gamba, P., Gaughan, A., Lisini, G., Stevens, F.R., Tatem, A.J., Trianni, G., 2015. Multitemporal settlement and population mapping from Landsat using Google Earth Engine. Int. J. Appl. Earth Obs. Geoinformation 35, Part B, 199–208. doi: 10.1016/j.jag.2014.09.005

Patel, N.N., Stevens, F.R., Huang, Z., Gaughan, A.E., Elyazar, I., Tatem, A.J., 2016. Improving Large Area Population Mapping Using Geotweet Densities: Improving Large Area Population Mapping Using Geotweet Densities. Trans. GIS. doi:10.1111/tgis.12214

Patterson, L., Urban, M., Myers, A., Bhaduri, B., Bright, E., Coleman, P., 2007. Assessing spatial and attribute errors in large national datasets for population distribution models: a case study of Philadelphia county schools. GeoJournal 69, 93–102. doi:10.1007/s10708-007-9099-3

Perkins, A.T., Siraj, A.S., Ruktanonchai, C.W., Kraemer, M.U.G., Tatem, A.J., 2016. Model-based projections of Zika virus infections in childbearing women in the Americas. Nat. Microbiol. 1, 16126. doi:10.1038/nmicrobiol.2016.126

Ruktanonchai, C.W., Ruktanonchai, N.W., Nove, A., Lopes, S., Pezzulo, C., Bosco, C., Alegana, V.A., Burgert, C.A., Ayikho, R., Charles, A.S., Lambert, N., Msechu, E., Kathini, E., Matthews, Z., Tatem, A.J., 2016. Equality in Maternal and Newborn Health: Modelling Geographic Disparities in Utilisation of Care in Five East African Countries. PLoS ONE 11. doi: 10.1371/journal.pone.0162006

Ruxton, G.D., 2006. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behav. Ecol. 17, 688–690. doi:10.1093/beheco/ark016

Say, L., Raine, R., 2007. A systematic review of inequalities in the use of maternal health care in developing countries: examining the scale of the problem and the importance of context. Bull. World Health Organ. 85, 812–819. doi:10.1590/S0042-96862007001000019

Schur, N., Herlimann, E., Traore, M.., Ndir, O., Ratard, R.., Tchuente, L.., Kristensen, T.., Utzinger, J., Vounatsou, P., 2011. Geostatistical model-based estimates of schistosomiasis prevalence among individuals aged <20 years in West Africa. PLoS Negl Trop Dis 5. doi: e1194

Soares, M.R.., Clements, A.C.A., 2011. Mapping the risk of anaemia in preschool-age children: the contribution of malnutrition, malaria and helminth infections in West Africa. PLoS Med 8. doi:e1000438.

Stevens, F.R., Gaughan, A.E., Linard, C., Tatem, A.J., 2015. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLoS ONE 10. doi: 10.1371/journal.pone.0107042

Tatem, A., Linard, C., 2011. Population mapping of poor countries. Nature 474, 36–36. doi:10.1038/474036d

Tatem, A.J., Adamo, S., Bharti, N., Burgert, C.R., Castro, M., Dorelien, A., Fink, G., Linard, C., John, M., Montana, L., Montgomery, M.R., Nelson, A., Noor, A.M., Pindolia, D., Yetman, G., Balk, D., 2012. Mapping populations at risk: improving spatial demographic data for infectious disease modeling and metric derivation. Popul. Health Metr. 10, 8. doi:10.1186/1478-7954-10-8

Tatem, A.J., Campbell, J., Guerra-Arias, M., de Bernis, L., Moran, A., Matthews, Z., 2014. Mapping for maternal and newborn health: the distributions of women of childbearing age, pregnancies and births. Int. J. Health Geogr. 13, 2. doi:10.1186/1476-072X-13-2

Tatem, A.J., Garcia, A.J., Snow, R.W., Noor, A.M., Gaughan, A.E., Gilbert, M., Linard, C., 2013. Millennium development health metrics: where do Africa's children and women of childbearing age live? Popul. Health Metr. 11, 11. doi:10.1186/1478-7954-11-11

Tatem, A.J., Noor, A.M., Hagen, C. von, Gregorio, A.D., Hay, S.I., 2007. High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa. PLOS ONE 2, e1298. doi: 10.1371/journal.pone.0001298

UNdata, 2016. UNdata | country profile | Mozambique [WWW Document]. URL http://data.un.org/CountryProfile.aspx?crName=Mozambique (accessed 10.17.16).

UNICEF, 2011. CHILD POVERTY AND DISPARITIES IN MOZAMBIQUE 2010 (Summary Report, UNICEF Mozambique). Maputo, Mozambique.

WHO, 2016a. WHO | Millennium Development Goals (MDGs) [WWW Document]. WHO. URL http://www.who.int/topics/millennium_development_goals/en/ (accessed 7.3.16).

WHO, 2016b. WHO | From MDGs to SDGs, WHO launches new report [WWW Document]. WHO. URL http://www.who.int/mediacentre/news/releases/2015/mdg-sdg-report/en/ (accessed 10.8.16).

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79–82. doi:10.3354/cr030079

Yang, X., Huang, Y., Dong, P., Jiang, D., Liu, H., 2009. An Updating System for the Gridded Population Database of China Based on Remote Sensing, GIS and Spatial Database Technologies. Sensors 9, 1128–1140. doi:10.3390/s90201128

ANNEXURE A: SEARCH TERMS

Methods (GIS/RS etc.)	**Population or demographics**
"Land use" **OR** "Land cover" **OR** "Remote sensing" **OR** "MODIS" **OR** "Pattern Decomposition" **OR** "Population Spatialization Model" **OR** "GlobCover" **OR** "Global land cover data" **OR** "Phone calls" **OR** "Mobile phones" **OR** "Ancillary layers" **OR** "Bayesian hierarchical spatio-temporal" **OR** "Bayesian hierarchical spatiotemporal" **OR** "night-time lights" **OR** "Landsat" **OR** "settlement maps" **OR** "satellite imagery" **OR** "Satellite derived maps" **OR** "Satellite maps" **OR** "Random forest" **OR** "covariate data" **OR** "HGPS" **OR** "Heuristic sampling" **OR** "QuickBird" **OR** "Frequency distribution function" **OR** "CART" **OR** "ancillary" **OR** "multivariate regression model" **OR** "power exponential decay model" **OR** "regionally-parameterized models" **OR** "geospatial covariate datasets" **OR** "surface modelling" **OR** "street weighting" **OR** "raster pixel map" **OR** "NLCD" **OR** "CORINE Land Cover 2000" **OR** "MDA GeoCover" **OR** "telecommunications data" **OR** "gravity model" **OR** "scale invariance" **OR** "cell phone network" **OR** "call detail record" **OR** "ethernet networks" **OR** "mobile computing" **OR** "pattern clustering" **OR** "space-time multiple regression model" **OR** "mobile calls" **OR** "Kriging" **OR** "building extraction" **OR** "image classification" **OR** "SIOSE" **OR** "Kernel width" **OR** "distance decay parameter" **OR** "areal weighting" **OR** "Dasymetric modelling" **OR** "Dasymetric" **OR** "Pycnophylactic" **OR** "intelligent dasymetric mapping" **OR** "binary dasymetric mapping" **OR** "Cadastral-based Expert Dasymetric System" **OR** "3-class dasymetric method" **OR** "multi-dimensional dasymetric modeling" **OR** "multi-layer multi-class dasymetric" **OR** "areal interpolation" **OR** "downscaling" **OR** "spatial disaggregation"	"population modelling " **OR** "gridded population " **OR** "Population distribution " **OR** "Settlement extent " **OR** "Census data " **OR** "Population mapping " **OR** "population densities " **OR** "population maps " **OR** "age-structure maps " **OR** "Spatial demography " **OR** "areal census counts " **OR** "demographic structures " **OR** "spatial covariates " **OR** "demography " **OR** "population estimates " **OR** "human mobility " **OR** "grid-based population " **OR** "geostatistics " **OR** "population surface " **OR** "population forecasts"