Digital Transport for Africa: Going Beyond Mapping in Kumasi

Final Report – Outline

Final Report v1

Date: January 20, 2023

Produced by:

Transport for Cairo مواصلة للقاهرة

For:

World Resources Institute

Digital Transport for Africa

Title picture: Virtual Capacity Building sessions with KMA (30/9/'22 WRI)

Table of Contents

List of Figures

Glossary

Acronym	Term
GTFS	General Transit Feed Specification
AFD	French Development Agency
WRI	World Resource Institute
TfC	Transport for Cairo
DT4A	Digital Transport for Africa
OSM	OpenStreetMap
GIS	Geographic Information Systems
PT	Public Transport
GKMA	Greater Kumasi Metropolitan Assembly
BRT	Bus Rapid Transit
SUMP	Sustainable Urban Mobility Plan
ERD	Entity Relationship Diagram
CSV	Comma-separated values

1 Executive Summary

This report explains the technical assistance provided to Kumasi Metropolitan Assembly's Department of transport by Transport for Cairo (TfC) under the supervision of World Resource Institute (WRI). The report details why the city of Kumasi was chosen to receive the assistance, and the activities done.

Well-defined objectives for the technical assessment were based on engagement with Kumasi Metropolitan Assembly's (KMA) Department of Transport. This resulted in conducting two main activities:

1. Data transformation and integration from different sources into KMA's existing data workflows
2. Capacity building for KMA's department of transport team to perform digital transit data collection

For activity (1) TfC sourced raw datasets from previous consultations and studies provided to KMA and transformed this data into a system currently utilized by KMA called "Transit Info Map". TfC transformed, cleaned, and integrated the raw datasets within the system. The extracted spatial dataset comprises the start and end points for 534 routes, the itineraries for 145 of those routes, and a digitized layer of the GKMA administrative subregions (i.e. submetros) . Additionally, there were some recommendations to adjust the Transit Info Map database schema to enable more features for users.

For activity (2) the KMA team were given access to TfC's developed Field Research Software suite "RouteLab" and given a full training on how to collect data on paratransit services by doing identification, onboard and frequency surveys. The training was done over 5 sessions during 2 consecutive days.

Finally, the report concludes with lessons learned from the exercise and what could be done in the near future. The main lesson learned came from the existing data assessment exercise, which was: Even though GTFS is a standard, that does not mean GTFS feeds are of equal quality and readily usable for analysis. GTFS data has to be properly assessed for quality before assuming it ready for decision making. This is further explained by introducing the Data Maturity Rating System, which was fine-tuned because of this exercise.

2 Background & Objectives

2.1 Kumasi's "Beyond Mapping"

A "Beyond Mapping" activity is defined as any activity which leverages existing geospatial and temporal datasets to conduct higher-level analytical, decision-making, or otherwise insight-generating activities. The primary goal is to drive impact through data based on each city context.

2.2 City Selection Methodology

The city of Kumasi was selected for "Beyond Mapping" technical assistance due to several factors:

1. It is a **key city** in WRI's prior projects, such as its Urban Road Safety Program launched in May of 2021. The presence of **prior successful collaborations** was noted as a positive entry point.

2. The city of Kumasi had a feasibility study conducted, in January of 2020, by ROM-Transport[1] and funded by AFD, in January of 2020, to ascertain the effectiveness of a **BRT intervention**. This project involved a data collection activity which provided the city with raw data for the city's transit routes. The data was also converted into a GTFS feed.
3. The presence of a GTFS feed with no discernible applications presented an **opportunity** to leverage this data for higher-quality, data-informed decision making.
4. The city's interest to leverage their existing data in further digitalizing the transit system by migrating it into their custom transit management tool **Transit Info Map**.
5. MobiliseYourCity Africa is currently conducting supporting activities[2] for the creation of a **Sustainable Urban Mobility Plan (SUMP)** for the Kumasi Metropolitan Assembly (KMA).

Once Kumasi was shortlisted, TfC and WRI reached out to key stakeholders at the Department of Transportation within the Kumasi Metropolitan Assembly in May 2022. Following the introductory call, a "Diagnostic Assessment Checklist" developed by WRI was sent to the local counterpart to obtain comprehensive details about the city's highest priority needs.

2.3 Project Aims

Upon reviewing the Diagnostic Assessment, it was apparent that KMA Transport Department had a clear direction and plan for their urban mobility future through the development of a system called **Transit Info Map.** Hence, the support activities are tailored to best support the municipality in their goals with Transit Info Map. These goals can be summarized as follows:

1. Leverage existing transit data to enrich the Transit Info Map system.
2. Clarify, with the help of the consultant, TfC, future steps to enhance both the data and the Transit Info Map system.

[1] https://www.rom-transport.com/_files/ugd/fcb6a2_a34c4f0325ed4df984f148ef8931db34.pdf?index=true

[2] Mobilise Your City Global Monitor 2022, https://www.mobiliseyourcity.net/factsheet-kumasi

2.4 Greater Kumasi Metropolitan Area

The Kumasi Metropolitan Assembly began as an area spanning a 10-15 km radius from the city center. It consisted of four geographic areas, known as sub-metropolitans or submetros. The Kumasi Metropolitan Assembly was officially granted Municipality status in 2018 and the geographic area (now spanning 40-60 kms from the center, having grown to nine submetros) is now regarded as the Greater Kumasi Metropolitan Area (GKMA)[3]. A raster sub-metropolitan map of GKMA is shown in Figure 1: Raster Map of GKMA Submetros (Source: Town and Country Planning Department, 2008). For the purposes of this report, any reference to KMA or GKMA is interchangeable and indicates the municipality which governs the following nine submetros:

1. Asawasi
2. Asokwa
3. Bantama
4. Kwadaso
5. Manhyia
6. Nhyiaeso
7. Oforikrom
8. Suame
9. Subin
10. Tafo

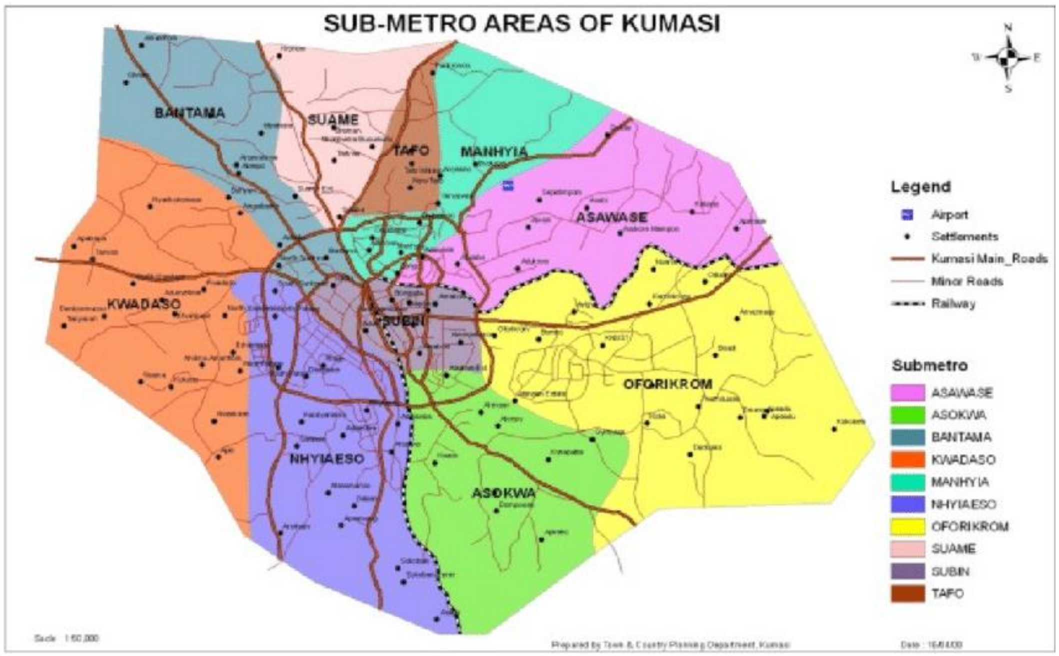

Figure 1: Raster Map of GKMA Submetros (Source: Town and Country Planning Department, 2008)

2.5 What is Transit Info Map?

This system is developed by Innoverex IT Solutions for the KMA Transport Department. The project began in 2017 following a project proposed by the Manhyia sub-metro for conducting a terminal identification study. The sub-metro needed to identify all informal terminals to gain a better understanding for later provisions to the public transport system. The submetro authority also needed to participate in regulating and operating these terminals to ensure higher-quality safety standards are met, and that competition with existing formal stations is reduced. Back then, the sub-metro had identified 107 stops and terminals of varying sizes, most of which existed without approvals from the

[3] Kumasi Metropolitan Assembly Service Charter, http://kma.gov.gh/kma_metro/docs/1040Service%20Charter_KMA.pdf

KMA. The effort conducted aimed to produce a well-labeled map with every terminal mapped out, and furthermore, linked to their corresponding routes.

Transit Info Map is a system with three components:

1. A **web-based dashboard** called "Transport Manager": designed to be used by the KMA and all sub-metro authorities for tracking and planning public transit in the city.
2. A **passenger-facing Mobile application**: used for trip-planning purposes
3. An **operator-facing Mobile application**: used by unions to monitor the drivers within their fleet and ensure compliance

To date, the dashboard and operator application are both live and currently being used by KMA. A manual data-collection effort was carried out to survey drivers and collect the necessary information to register them to the operator mobile application. The survey used is visualized below.

KUMASI METROPOLITAN ASSEMBLY
DEPARTMENT OF TRANSPORT / GIFEC / BIGRS-TRANSPORTINFOMAP
DATA COLLECTION FORM FOR TRANPORT OPERATORS-BRANCH (DRIVER/VEHICLE INFORMATION)

NB: To be filled or completed by any of the Terminal or Station Executives

DRIVER'S INFORMATION			
Sub-Metro/Terminal		Place of Issue	
Union		Date of Issue	
Route		Expiry Date	
Name of Driver		Vehicle No.	
Alias(Popular)		Vehicle Type	
Home Address		Brand and Model	
Digital Address		Capacity	
Tel. Number		Name of Owner	
Date of Birth		Contact Number	
License Number		Home Address	
Class of License		Insurance Company	

NB: Attach a copy of driver's license to the form when completed

Figure 2: Paper Survey for Driver Registration (Source: Department of Transport, Kumasi Metropolitan Assembly)

2.6 Stakeholder Mapping

- **Digital Transport for Africa (DT4A) Network:** is an initiative funded by Agence Française de Développement (AFD) and led by the World Resources Institute (WRI). DT4A aims to create a movement for open transit data across the African region. DT4A represents a diverse network of city governments, civic technology companies, collectives, residents, universities and international development organizations. The network is now scaling up efforts in data collection, mapping, analysis, and capacity building with the aim of reaching 10 additional cities across Africa.

The DT4A resource sharing platform is also moving beyond mapping just data on public transport in cities, but seeks to inform planning, creation of tools, and meaningful research on actions to improve sustainable mobility.

- **French Development Agency (AFD):** is a public financial institution that implements the policy defined by the French Government. It works to promote sustainable development and fight poverty. AFD Group finances and supports development projects in; energy, urban development, education, climate, agriculture, health, water and sanitation, digital technologies, biodiversity and sport. AFD is one of the main sponsors of the DT4A initiative.
- **World Resource Institute (WRI):** is a global research non-profit organization that works on tackling multiple global challenges including mobility in cities. WRI-Africa is currently leading the DT4A project implementation and have hired Transport for Cairo as technical support consultant
- **Transport for Cairo (TfC):** TfC provides data, tools, and research to improve urban mobility in emerging cities, primarily in Africa. They are the technical support consultant responsible of the project implementation. Namely to undertake mapping and going beyond mapping projects in Alexandria, Egypt; Kumasi, Ghana and Maputo, Mozambique.
- **Department of Transport, Kumasi Metropolitan Assembly:** Responsible for transport regulations and planning within Kumasi city, and main beneficiary of the project.

As mentioned in 2.3, the municipality aims to unite transportation in its entire area of control under the Transit Info Map digital system. Though the submetro of Asawase may not fall under the jurisdiction of the KMA authority, for all intents and purposes it will be included in this report as the region's transportation needs impact the GKMA.

Aiding the Department of Transportation are the following entities which will also benefit from these "Beyond Mapping" activities:

- KMA Physical Planning Department
- KMA Planning Department
- Department of Planning, Kwame Nkrumah University of Science and Technology
- Other Municipal Assemblies in Greater Kumasi

2.7 Defined Project Objectives

This section provides a list of the objectives that guided the project's activities. These objectives are based on both the initial goals of the project, explained in 2.3, and the outcomes of the assessment phase, explained in 3. The objectives are as follows:

1. Conduct a **comprehensive Data Assessment** on pre-existing datasets to identify strengths, weaknesses, and opportunities for integration with Transit Info Map.
2. Provide **capacity building** to members of Department of Transportation, Physical Planning Department, and others to improve their GIS expertise within these organizations, specifically in the context of transport data collection and management.

- *Note: As clarified by the counterpart, given that the level of training required for all self-identified areas of improvement are equal (Diagnostic Assessment Checklist), TfC and KMA jointly chose to focus on (i)* ***Project Planning & Management*** *and (ii)* ***Data Collection & Analysis.***

3. Conduct a **guided data transformation** activity whereby existing datasets are transformed into file formats and data structures that are compatible with Transit Info Map. This is to primarily enable the consumption of GIS data through the more simplified and user-friendly interface of Transit Info Map over the use of technical tools such as QGIS or ArcGIS.
4. Make the outputs of the data transformation activity openly available on the DT4A resource center's data repository.

3 Assessment Phase

We understand data systems as a pipeline which flows from Data collection all the way to Impact Mechanisms (such as Transit Info Map):

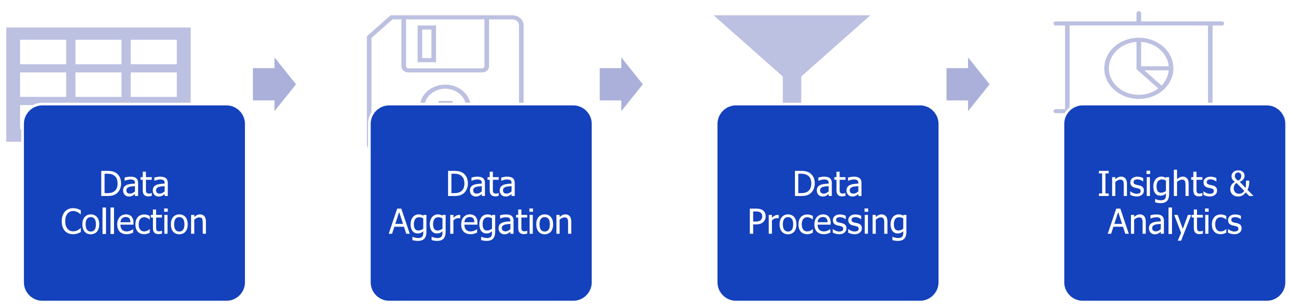

Figure 3: Data Pipeline Diagram

In the case of the KMA:

1. TfC is collaborating with the municipality (an Information Source with preexisting Raw Data Sources
2. Evaluate the existing model (GTFS Feed) and, where possible, provide a better model (the proprietary data schematic of Transit Info Map)
3. Aid the municipality in best consuming their available data through a Dashboard (provided by Transit Info Map).

3.1 Assessing Existing Data

TfC, through its transit data collection and analysis experience, has developed a **Data Quality Model** for assessing the quality of transit data. The model consists of a set of data points that range from basic to advanced, based on the data collection effort involved, and cover geographic, temporal and system adequacy aspects of a transit system. A detailed explanation of the model is provided in Appendix A: Benchmarking: Three-Pillar Model.

The existing transit datasets for GKMA are the following:

1. Raw data collected by ROM-Transport in 2020
2. GTFS feed created based on raw data in the previous point

In Appendix B: Existing Data Assessment, an extensive assessment of the datasets, based on the Data Quality Model mentioned above.

Said assessment showed the following:

1. The datasets cover basic geographic data points (trip geometry, stop locations and stop names). A lack of temporal data, such as headway and time between stops, was observed.
2. A lot of duplicate and unused stops were found in the GTFS.

TfC decided, based on the assessment, to utilize the raw data directly in the projects data-related tasks, defined in 3.3.

3.2 Assessing "Transit Info Map"

TfC identified a critical missing component in the Transit Info Map software. The applications were designed on a database which did not support geographic entities. As a result, while the application contained tables of all categorical and numeric data, it lacked any geographic visualizations due to the lack of support in storing such data.

Thankfully, Innoverex had already identified this gap in the system and support for geographic entities is within the product roadmap. TfC identified this as a potential for collaboration or capacity building, though outside the scope of this project.

3.3 Defining the Activities

3.3.1 Leveraging Existing Raw Data

The existing raw data and GTFS feed, although accessible to KMA, were not utilized in the Transit Info Map system currently in use. The database schema of Transit Info Map is not compatible with the GTFS format. Therefore, we helped KMA leverage the raw data they had by building a pipeline that cleaned and transformed the raw data into inputs directly compatible with the Transit Info Map schema.

3.3.2 Capacity Building

Given the lacking nature in the available datasets—from the missing temporal dimension to the very limited number of available data points—the number of potential applications is consequently limited.

Initially, we believed the best application is to provide capacity building on transit data collection, based on the three-pillar model explained in 2.4, as well as the types of analyses that may be conducted with such data, such that it may be utilized when the data is made available.

After examining the data and consulting with the beneficiary, we identified that the city has already started consolidating their data in Transit Info Map. DT4A and TfC's approach was to provide value regardless of ownership of a given solution. The team decided on helping the city integrate datasets from data collection done by consultants, that was previously shelved, into the Transit Info Map system to have a more comprehensive vision of the city's network through the system.

4 Activity 1 – Raw Data Transformation

4.1 Introduction

The Transit Info Map data management system uses a custom database schema. An Entity Relationship Diagram (ERD) was received from the client. The purpose of this activity is to convert the existing raw data, collected by ROM Transport Engineering, into outputs that can be either fed directly into the Transit Info Map database **or** enrich the system's data by informing manual processes to create the data.

4.2 Pipeline Inputs and Outputs

4.2.1 Inputs

The raw data input was a set of "routes", each with a start point, an end point and an itinerary in the form of tabular points in a CSV file, as shown in Figure 4 and Figure 5.

ID	Route	Link End	Link Car Route	ID	Start Point	End Point
36	Abuakwa_ AdumUnibank	https://www	https://www.google.co.il/maps/dir/6.700420,-1.715013/6.694936,-1.623245	36	6.700420,-1.715013	6.694936,-1.623245
37	DrMensah_ OpokuTradingAdum	https://www	https://www.google.co.il/maps/dir/6.699287,-1.619647/6.690226,-1.623472	37	6.699287,-1.619647	6.690226,-1.623472
38	OpokuTradingAdum_ DrMensah	https://www	https://www.google.co.il/maps/dir/6.690286,-1.623562/6.699352,-1.619240	38	6.690286,-1.623562	6.699352,-1.619240
39	Manhyia_ Pampaso Station	https://www	https://www.google.co.il/maps/dir/6.704997,-1.720140/6.697102,-1.624307	39	6.704997,-1.720140	6.697102,-1.624307
40	Pokukrom_ Kejetia CBG	https://www	https://www.google.co.il/maps/dir/6.719642,-1.700844/6.697195,-1.622940	40	6.719642,-1.700844	6.697195,-1.622940

Figure 4: Route names with start and end locations

Longitude	Latitude	Altitude	Name	Start/End	
-1621009	6694483		AdumBossFm_Ejisu	Start	AdumBossFm_Ejisu-Start
-1621009	6694483	286.7	AdumBossFm_Ejisu		
-1620995	6694417	286.7	AdumBossFm_Ejisu		
-1620986	6694346	286.7	AdumBossFm_Ejisu		
-1620941	6694278	286.7	AdumBossFm_Ejisu		
-1620906	6694222	286.7	AdumBossFm_Ejisu		
-1620931	6694027	286.7	AdumBossFm_Ejisu		

Figure 5: Corresponding itinerary points for routes

4.2.2 Outputs

Of many entities in the Transit Info Map schema, the raw data could be transformed to cover a few basic entities; namely, routes, terminals and submetros.

Terminals, in TfC's definition, are locations or main hubs were routes start or end, meaning that a route must have an origin terminal and a destination terminal. But the Transit Info Map, as shown in Figure 6 is different. The relation between routes and terminals in Figure 6 is independent of a route's origin and destination.

Suspecting it is a technical issue in the schema design, the team asked the Innovorex team, who clarified that this design is deliberate. It is justified by an interpretation of the "terminal" other than TfC's which is closer to a "main stop". A place where buses either start or end, or almost always stop to load or offload passengers. Therefore, a set of terminals cannot be simply created from the raw data but require a manual process. Thus, we only extract the routes' start and end locations in spatial format to inform such a manual process.

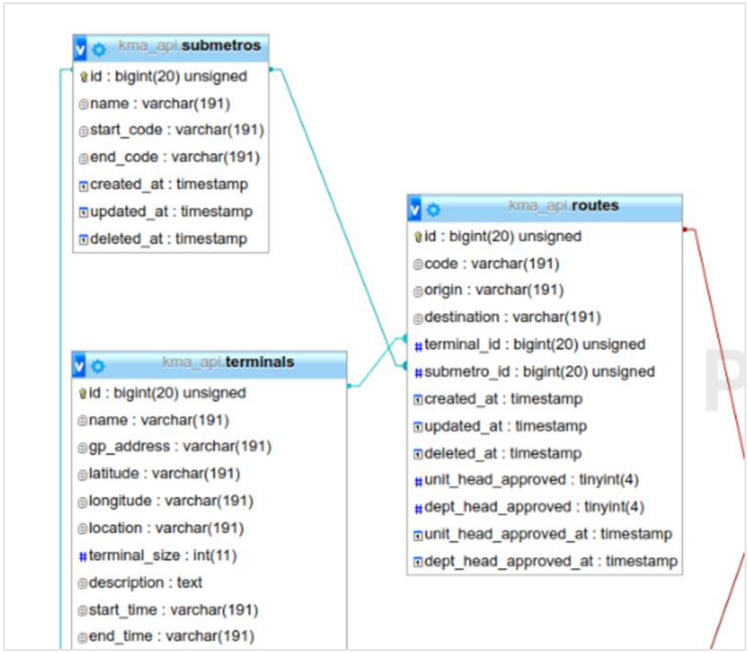

Figure 6: Subset of the Transit Info Map schema

The output of this pipeline comprises:

- Routes
  - Start and end locations (extracted from Excel files)
  - Raw itinerary linestring geometry (extracted from Excel files)
  - Itinerary geometry matched to OSM road network
- Digitized submetros

4.3 Steps

4.3.1 Digitization of Submetros

No spatial layer of the GKMA submetros was found online. We manually digitized them, using QGIS, based on a raster map provided by the client. They were in turn converted to both spatial GeoJSON, and tabular CSV files.

In the Transit Info Map Schema (see Figure 6), routes are related to the submetros. The raw data didn't have this information. However, we could leverage the spatial nature of the data. A route is simply associated with the submetro within which its origin is.

Processing Routes' Start and End Locations

The start and end points for the routes were observed to be completely raw inputs, where they lacked consistency in naming and coordinates. For example, the same place/landmark would be found as a start or end location for multiple routes but with slightly different coordinates and naming. Figure 7 demonstrates this observation with a subset of the raw points.

4.3.2

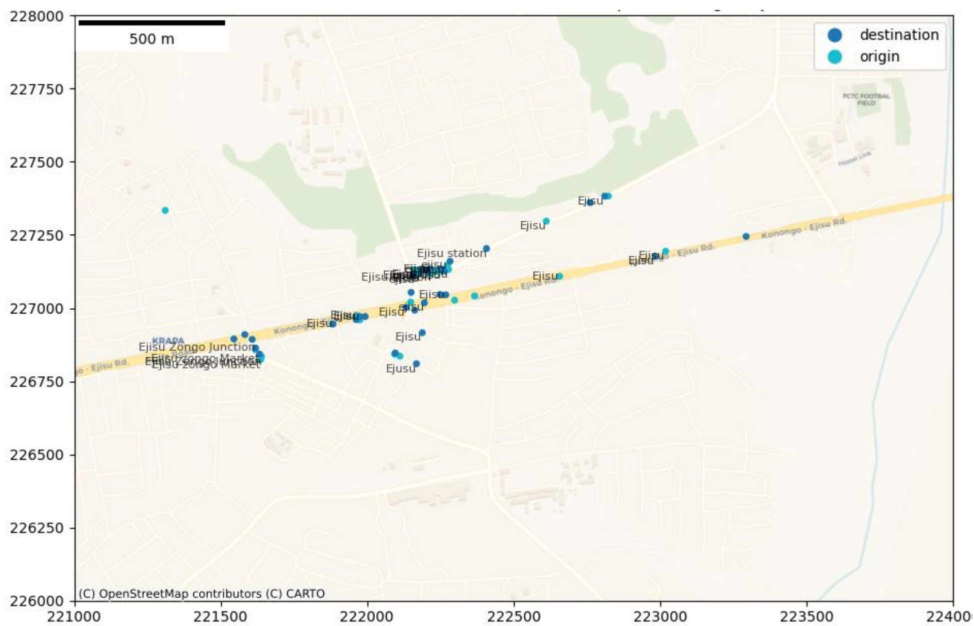

Figure 7 Subset of extracted points (Close-up of Konongo - Ejisu Road)

4.3.2.1 Processing steps

Detailed steps of the process are as follows:

1) **Extract raw start and end locations from the routes raw data**
   - **a)** Extract the points from the excel-sheet-based raw format into standard GeoJSON geometries.
     - **i)** 534 unique routes with complete data were found in the raw dataset. Their start and end points were extracted (1068 points)
2) **Spatially join locations to the submetros:**
   - a) Every location takes the \`submetro_id\` and \`submetro_name\` of the submetro that encloses it. Locations outside of Kumasi submetros take a \`submetro_id\` value of NULL
   - b) Another field is \`nearest_submetro_id\`. It is equivalent to \`submetro_id\` for locations within the Kumasi submetros. Each location outside Kumasi is assigned the id of the submetro that is spatially nearest to it.

4.3.2.2 Unsuccessful Processing Attempts

Even though we knew terminals would need to be manually created as mentioned in 4.2.2. As an attempt to shrink the gap between the raw points and terminals, we tried spatial clustering using both DBSCAN and HDBSCAN algorithms. But we concluded that such methods cannot produce clusters that accurately mirror the real terminals (seeing terminals as clusters of passenger activity). Local knowledge, along with consideration for the naming of the raw points are key to consolidating the points into representative set of terminals for the GKMA.

The following figures visualize the data outputs.

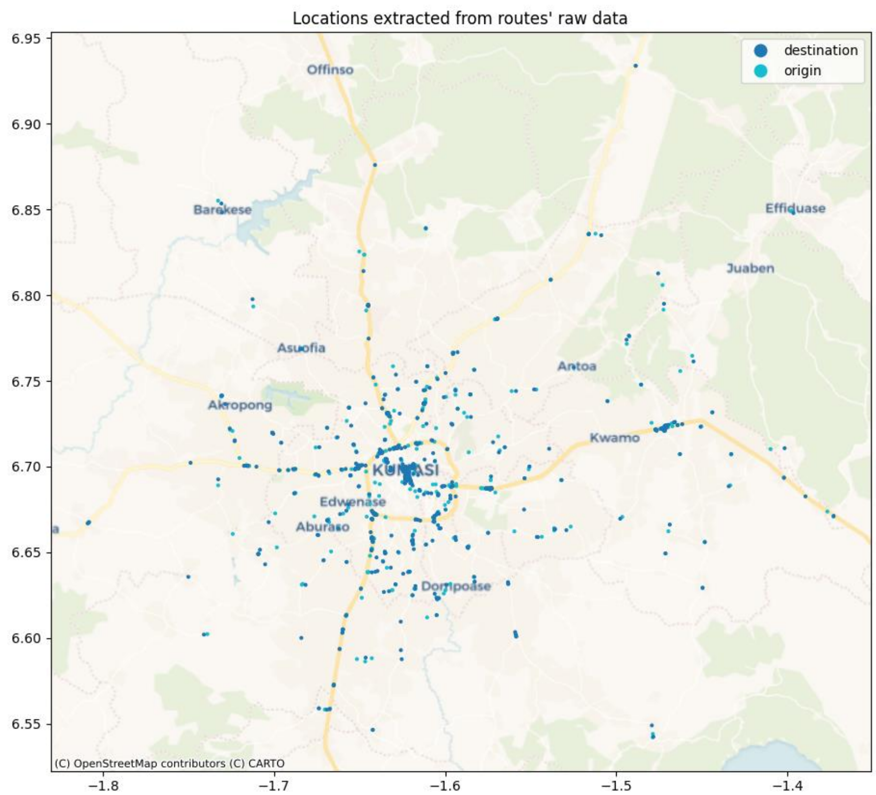

Figure 8: Raw extracted start/end points

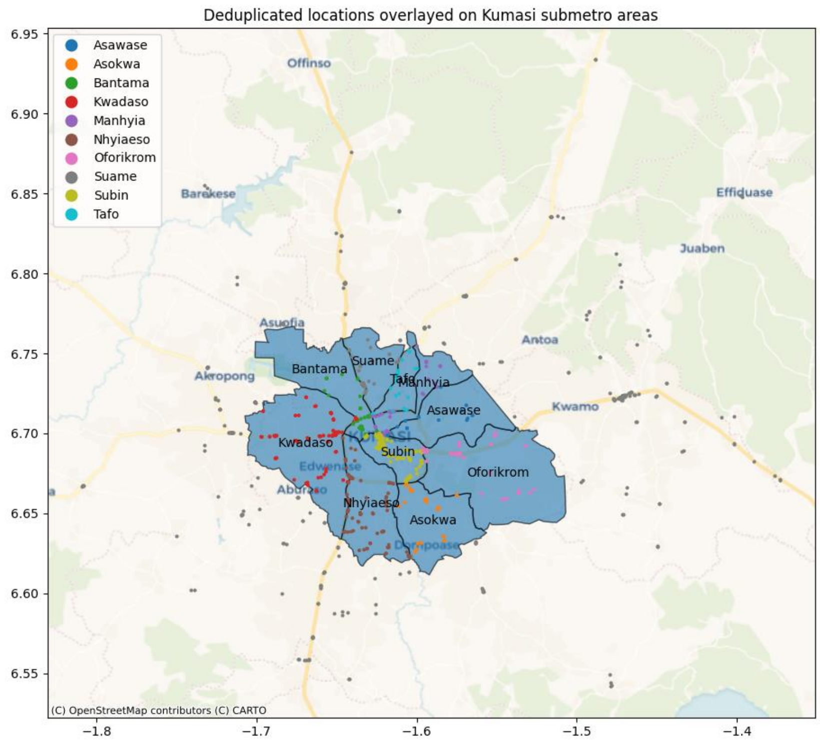

Figure 9: Processed route start/end points with GKMA submetros

4.3.3 Routes

A table of routes is extracted from the raw data along with the boarding and alighting locations from the previous section. For each route's origin and destination point, attributes for the spatially nearest submetro and unique location are attached.

4.3.3.1 Spatial data points processing

For only a subset of 145 routes out of 534, the route geometries were found in the raw data in the form of consecutive coordinates. were found in the raw data. Those were extracted into GeoJSON Linestring geometries.

These raw linestrings were found to not align with the road network, probably due to GPS logging errors. Using the route matching engine **GraphHopper**, these linestrings were matched to the OpenStreetMap (OSM) road network. The OSM road network was downloaded from BBBike.

4.3.3.2 Results

Resulting data is a CSV table containing 534 routes. 145 out of those are included in spatial layers, the raw linestrings and the linestrings matched to the OSM road network.

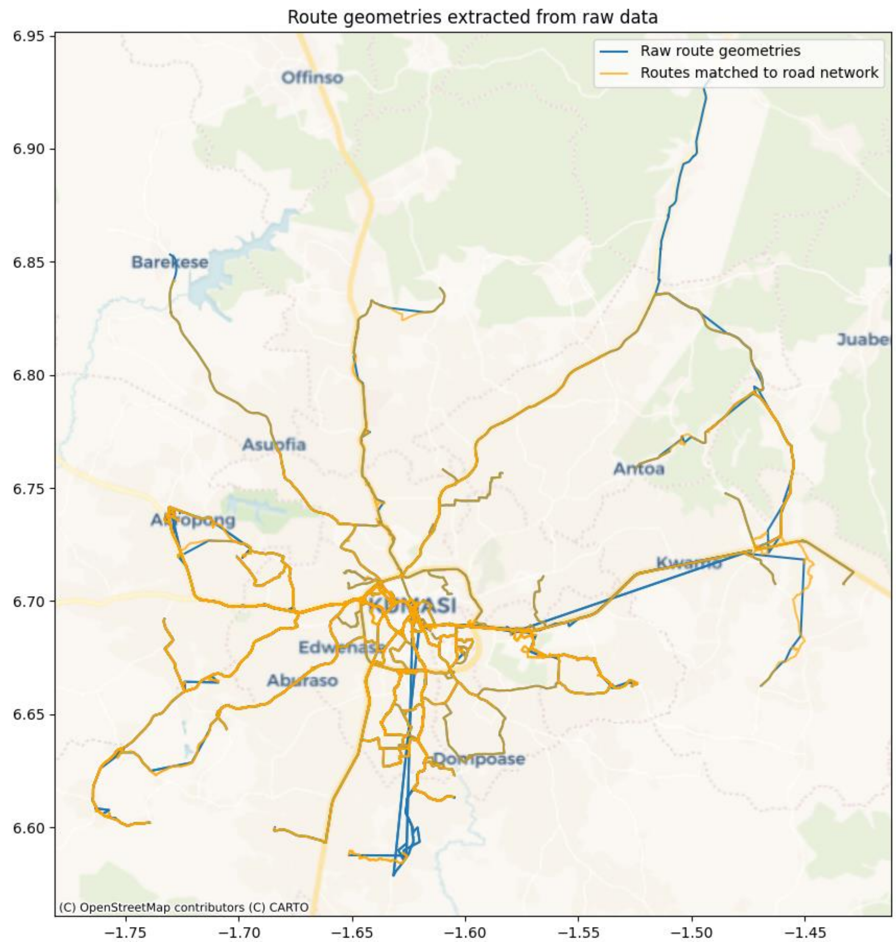

Figure 10: Raw and matched geometries of GKMA routes

4.4 Outputs

The main output dataset consists of the following tables in CSV format, with the required fields for the KMA schema:

- Routes
- Submetros

Further, the following spatial layers are also exported:

- Digitized submetros layer
- Routes' start and end points
- Routes (raw and matched geometries)

5 Activity 2 - Capacity Building

Second essential component of the services provided to the city of Kumasi was capacity building for digital data collection and applications in analysis and planning. The activity was coupled with the Transit

Info Map data integration exercise. The more operational hands-on of the former, and the more strategic long-term benefit of the latter.

From the first interview with the beneficiaries, a clear ask for a capacity building program—to enable local talent to conduct various geospatial analyses—was raised. This capacity building exercise would be used as a proof-of-concept rather than a viable analysis that can be used for policy-planning.

5.1 Agenda Specification

The design of the capacity building program meant to include different formats to keep the attendees engaged, namely:

- Lecture format: Where TfC presents virtually from Cairo with physical classroom attendance in Kumasi and physical facilitation by DT4A team
- Field surveying: Hands-on application of data collection by attendees using RouteLab
- Quiz and take-home assignment: specifically for the session about project management, attendees would receive an annotated version of their assignments with notes and feedback from the TfC team

To allow for more specialization, attendees into two groups for a portion of the training where each group takes a "deep dive" into one of two specialized topics: Field surveying & Survey project management.

Before agreeing on topics and the final agenda, there had to be back and forth communication with the beneficiary, KMA Department of Transport (DoT), to define their priorities. They clearly expressed the need to have ongoing data collection capacity, and also a preliminary understanding on "how" they can use this data for their benefit.

Given that input and the time availability of the city's employees and the duration of time WRI team would be physically present in Kumasi, the duration of the capacity building program was deemed to be two days.

Table 1: Detailed agenda for the capacity building program

Day	Session Title	Objectives
1	Digitizing Informal Transit – First Steps & Project Planning	• What is Data in Transportation? • What is the Paratransit Context? • What is the composition of a typical Data Collection project? • What are the main elements and process of mapping paratransit services? • What are some of the challenges and risks involved? • How do we estimate the cost of a Data Collection project?
	Break	

	Deep Dive Track #1: Uncovering Data & Technological Challenges	• What is a Transit Information System? • What types of data do we collect? • How do we go about collecting each data point? • What is an Identification, Onboard, & Frequency Survey? • What is Road-Side Interviews? • How does each survey contribute to our understanding of the network? • What are some Open-Source tools for Data Collection? • What is the advantage of using RouteLab? • RouteLab Demo
	Deep Dive Track #2: The Logistics & Operations of Data Collection	• How do we determine an adequate sample size given budgetary constraints? • What is the highest priority/highest impact Terminals & Routes to collect? • How do we setup a data collection timeline? • What are external forces/actors, and how do we account for them? • How do we recruit and train field researchers? • Setting expectations and reacting to deviations in the field • What KPIs do we monitor to ensure a successful data collection effort? • Key learnings from managing a large team
2	Hands-on Data Collection with RouteLab	
	Break	
	Beyond Mapping: Data Exploration, Mining, and Potential	• Why do we need a processing pipeline? • Identifying & removing anomalous data points • What is Graph Hopper, OSRM and how are they used to clean GPS traces? • How do we calculate headways? • Packaging GIS layers into a GTFS feed • Use-case #1: Job Accessibility Analysis • Use-case #2: Supply & Demand assessment • Use-case #3: Passenger Information Systems

	• Use-case #4: Journey Planning applications • Use-case #5: Vehicle counts with machine learning

The sessions were attended by approximately 26 individuals from the organizations highlighted in section Stakeholder Mapping2.6; this is in addition to 4 members from WRI who supported the success of the sessions on the ground.

5.2 Session 1: Digitizing Informal Transit

First session is an introductory session to the topic of digital data collection. The session was a lecture addressing two topics: "What do we mean by data in transport?" and "Field research in project management terms". The content was kept at a basic-intermediate level of depth to keep all attendees from the different backgrounds engaged and to unify the understanding of the topic.

Explaining "data in transport" meant we define quantitative data and how it's used to generate information and transform that into knowledge, The trainers explained the GIS data formats (raster & vector) and how those are used to represent the real world. There was also a special focus on the GTFS format, its history and applications, to highlight its importance as a dataset in a typical transport planning inventory.

Once the technical definitions of data are explained, the trainers described the characteristics of paratransit modes and how their special nature affects the methodology of collecting data and estimating parameters. For example, the modes are mostly demand-driven which means they need proper temporal sampling to make sure the field survey captures the discrepancies in service levels across the day.

The second part looked at data collection as a project and how it can be planned for and managed. TfC listed the standard activities and their sequence (i.e., Methodology design, obtaining permits, mobilization, and data cleaning), the parameters to take into consideration when planning (i.e., survey duration, sample size, team size, etc.) and how those parameters help transit agencies and municipalities estimate budgets for data collection and mapping projects.

Lastly, TfC detailed the special challenges faced in such projects based on previous hands-on implementation in different cities across Africa. This went into depth, for example TfC explained the weather conditions challenge in cities near the equator where rainy seasons hinder the transport services, distort GPS signal, and weaken internet connectivity, or how poor cellular network coverage in the country dictates choosing a network operator carefully when buying sim cards for the field surveyor phones.

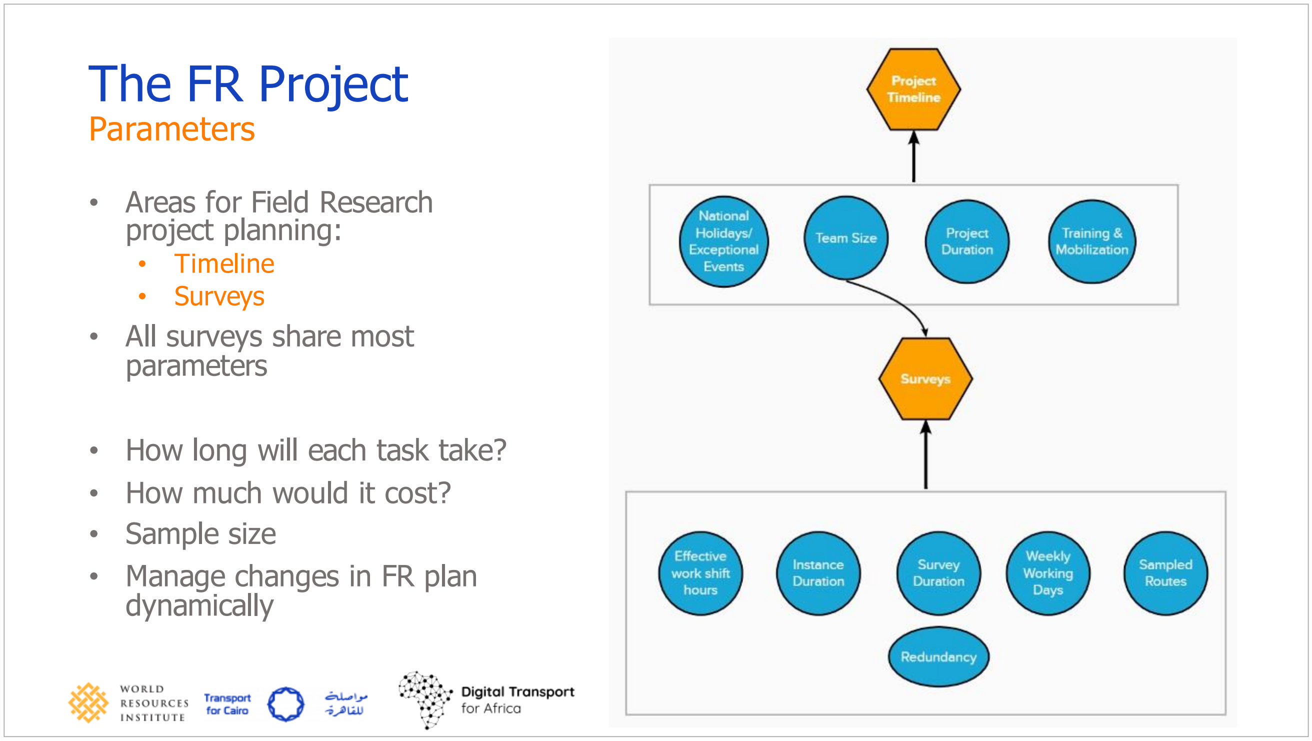

Figure 11: Excerpt from Session 1

5.3 Deep Dive Session 2a: Data & Tech Challenges

The session presented the unique challenges of collecting digital data for paratransit. The session started with detailing the different dimensions of digital data for any given transport network (Geographic, Temporal, and System Adequacy mapping) and going into the different levels of maturity (basic, intermediate, advanced) as shown in Figure 12.

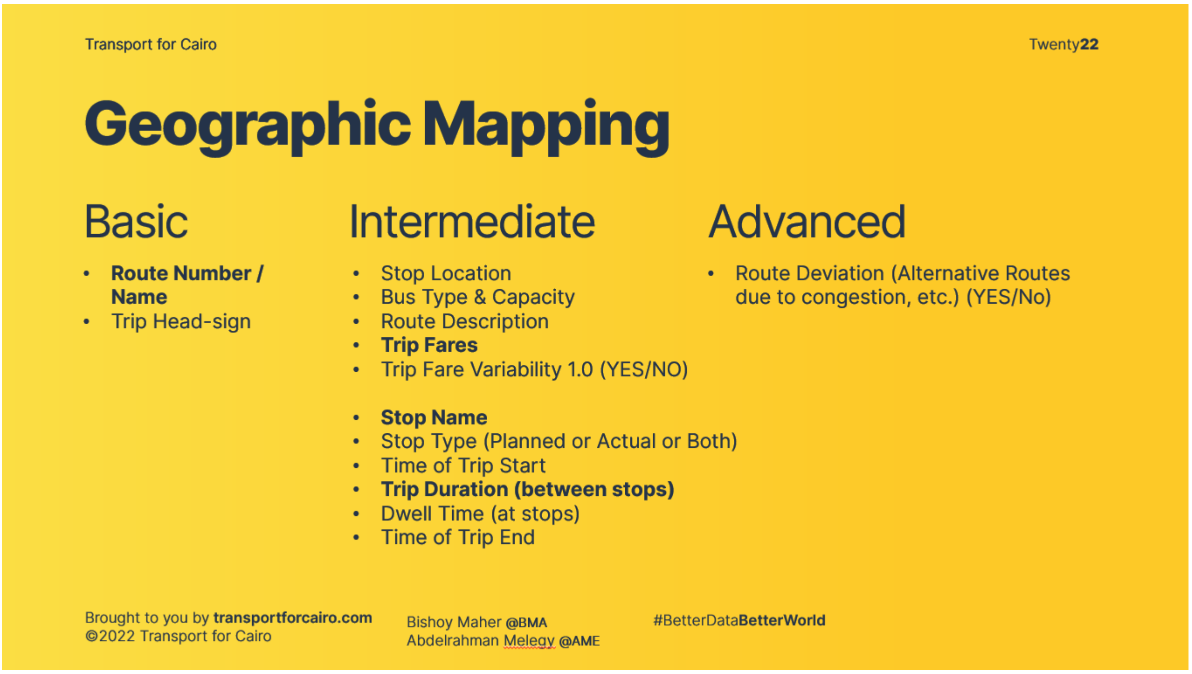

Figure 12: Excerpt from Session 2

The problem statement was concluded by displaying the current methods of data collection and their shortcomings. The session then transitioned into the application of RouteLab in a sequential manner, from the point of setting up a data collection project on the platform to data collection and cleaning. The presenter detailed each module's features along the way to show how it addresses a specific pain point in the data collection process.

The session ended with a hands-on demo on the RouteLab dashboard where each attendee was given credentials to access the web user interface and a link to download the "Observer" app to get a feel of how the software suite links field surveys done through the app with the management back office through the dashboard and lastly the GIS specialists working on data processing and validation of incoming surveys.

5.4 Hands-on Data Collection

In preparation for the data collection demo, an identification activity was scheduled first. This process, initially covered in depth in Session 2a, was followed by hands-on support to aid KMA's team in learning the tools and visualizing terminal polygons on the map for the first time. The identification process concluded with the digitization of 11 major terminals around the venue, and one dummy terminal at the venue, in which the training was held. Following terminal digitization, routes were also identified and created on RouteLab. From there, the KMA team was ready to enter the field.

On the second day of training, a set of researchers from the KMA team took to the streets to conduct Onboard Surveys. Given the constrained nature of the training schedule TfC opted to exclude Frequency Surveys from the demo, though the full process was covered in Session 2a.

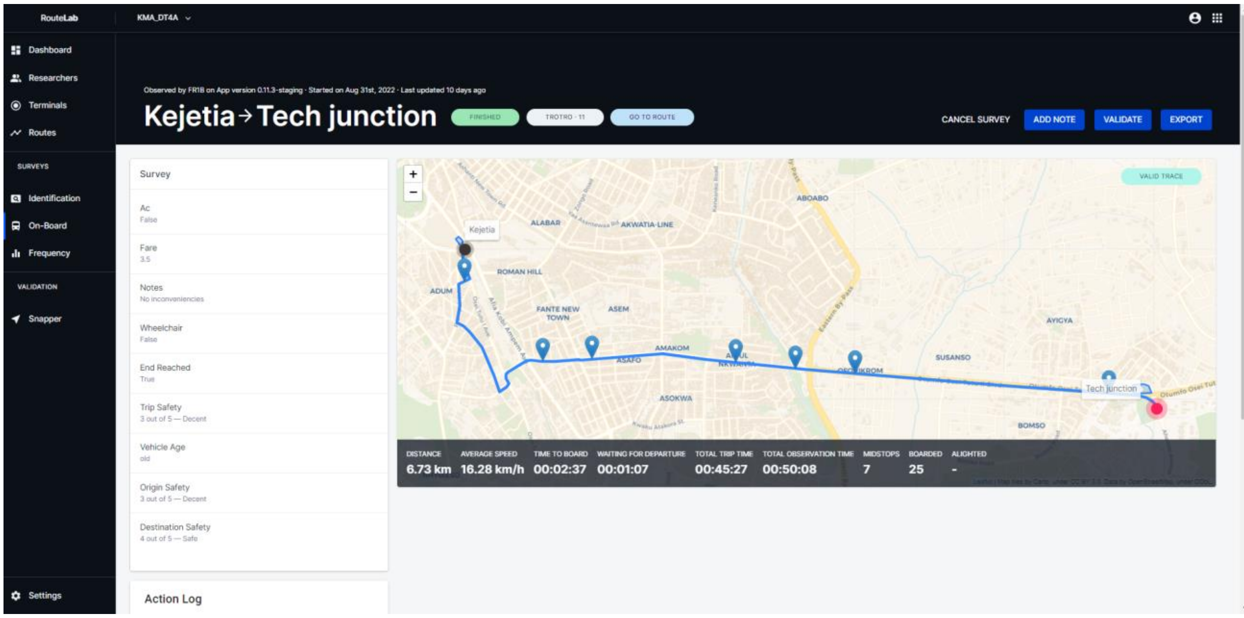

Figure 13: Onboard Survey from Kejetia to Tech Junction station with Boarding & Alighting locations

In total, the KMA team conducted **13 Onboard surveys** across **6 Routes** (i.e.12 route-legs). Every route-leg was surveyed at least once in this demonstration.

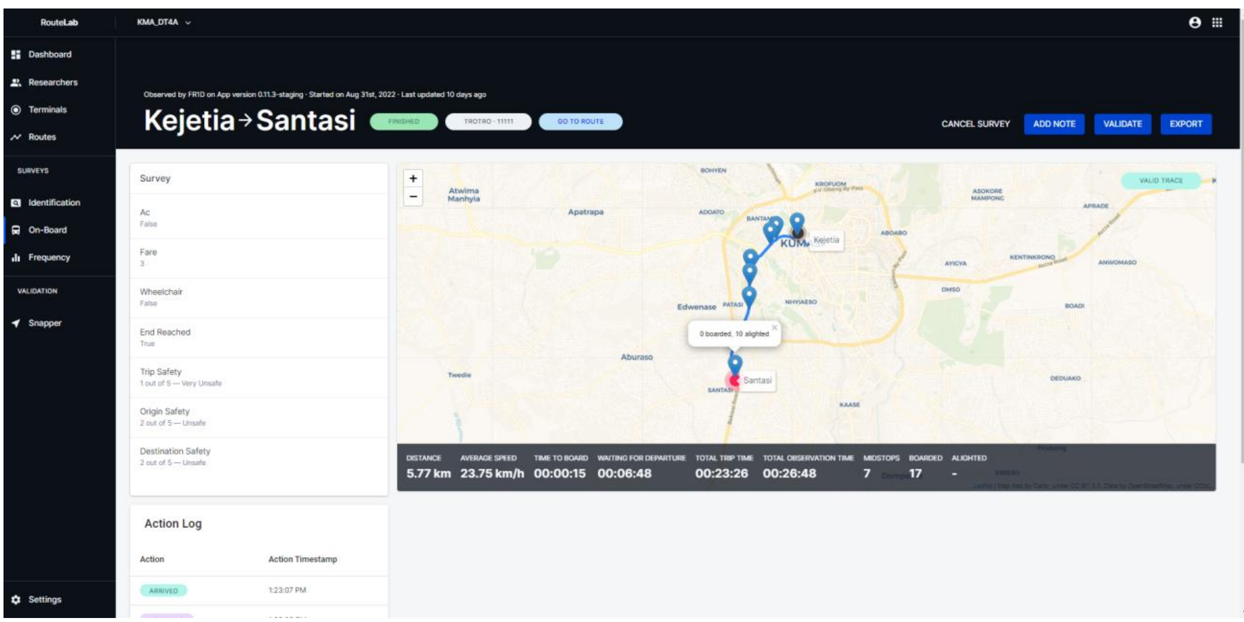

Figure 14: Onboard survey from Kejetia to Santasi with Boarding & Alighting locations

The data also included boarding and alighting data; both location, and number of passengers embarking & disembarking the vehicle at each stop. RouteLab also captures trip metrics such as Passenger Wait Time (time to board + time to depart), as well as the total Trip Time and the Total Number of Passengers Served.

Though this mini-data-collection effort is not comprehensive enough to generate any useful insights about the network, it does however demonstrate the power of RouteLab to the KMA team which should inspire and encourage them to continue using it for the duration of the project. Though a formal Data collection effort was not part of the scope of this project, the KMA team had access to RouteLab for the full duration of the project and were also given channels for support should they require any further assistance with the tools. However, it would be the KMA's responsibility to continue to collect data after the capacity building program concluded.

5.5 Deep Dive Session 2b: The Logistics & Operations of Data Collection

The session addressed attendees from middle and upper management, where the main topic is how to plan and manage a field research project. The first topic addressed was the sampling strategies, approaches and how to properly prioritize resource allocation (e.g. defining an index for "route importance", creating sampling units or traffic analysis zones "TAZ").

We then moved into the topic of the "Field Researchers". How should the team be composed? How are responsibilities divided amongst the different roles?

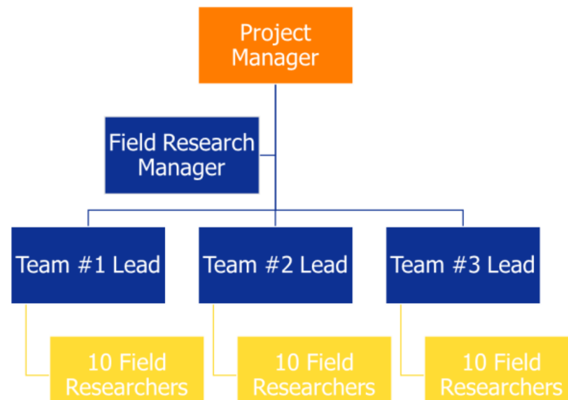

Figure 15: Example of a Standard Field Research Team Structure

The job profile and required skills for field researchers were explained in detail, what are the must-have skills (e.g. public transport user, local knowledge of the city) and what would be preferable to have (e.g. education background in geography or social science, experience with digital maps). There was a focus on the OSM community and their unique competence of accomplishing such tasks.

Then there was a deep dive into the project management aspects of data collection. We explained the activities, how to construct a timeline where each task is defined and assigned a duration and resources. We also explained the potential risks and how to mitigate them, based on previous experience from TfC and other consultants. Those risks range from project design and agility in responding to changes in design, to accounting for time required in the field, to security aspects and how to ensure safety of human and other resources in the field.

Finally, there was a group activity where attendees were divided into groups and asked to outline an abstract plan for a field research project. The team would do this outline over Word documents and deliver the output to the TfC team, where TfC would get back to individual groups with notes and feedback.

Group Activity excerpt from Session 2b:

- In groups of 4 or 5, we are to plan a paratransit mapping exercise from start to finish
    1. List the surveys, activities and timeline for the project
    2. Budget for software and hardware expenses and team members fees

3. Mention expected challenges to be faced, especially in due to local context

- We can use either digital tools such as excel, word, etc. (preferable) or draft on a piece of paper and scan to share with the rest of the team
- After each exercise we'll pause and give time for each time to share and review

The outputs from this exercise were quite satisfactory to the TfC team as it showed an understanding of the topic and the much-required contextual input from each team showing local understanding of their own challenges as well.

Timeline

The project will done in two (2) working days with six (6) hours each day and two (2) hours of break in between. In bound thus from an origin to a destination will be three (3) hours then a two (2) hour break followed by another three (3) hours of out bound thus from destination terminal to origin terminal for each route on two (2) separate days. – Are those inter-city routes? 3 hours seems a bit too long for a standard route within a city. It's very good how you broke down the tasks into hours and summed that up into working days

Team composition

The team shall comprise of

- A field research manager,
- Two team leaders and
- Ten field researches

Team criteria – Very good. It's good to define which requirements are for which team members (i.e. field researchers requirements are not the same as data analysts for example)

Figure 16: Snippet from one of the group assignments with TfC's comments in red

5.6 Session 3: Beyond Mapping

The last session meant to wrap things up with the question asked initially by the beneficiary: "How can we use this data to our advantage?". The session took a step back to look again at the bigger picture and gave an overview on the context of developing cities and their growth rate and pose the question of how we're properly going to meet the growing mobility demand.

A framework for addressing this issue is coined simply as "Map, model, transform" where:

- **Map**: Digitalization of transport data and maintaining it. This includes field data collection but also digitization of data existing in physical format
- **Model**: Aggregate, model and analyze data to come up with insights about specific real-world questions (e.g. where is the over-supply in a given city's corridors happening and when? Where

do people face problems of accessibility in their daily commute? What are the causes of those problems?)

- **Transform**: Take informed decisions based on the knowledge created from said data. An example is given for the case of Cairo's ring road BRT which is currently ongoing construction and how data contributed to the corridor selection and prioritization of investment

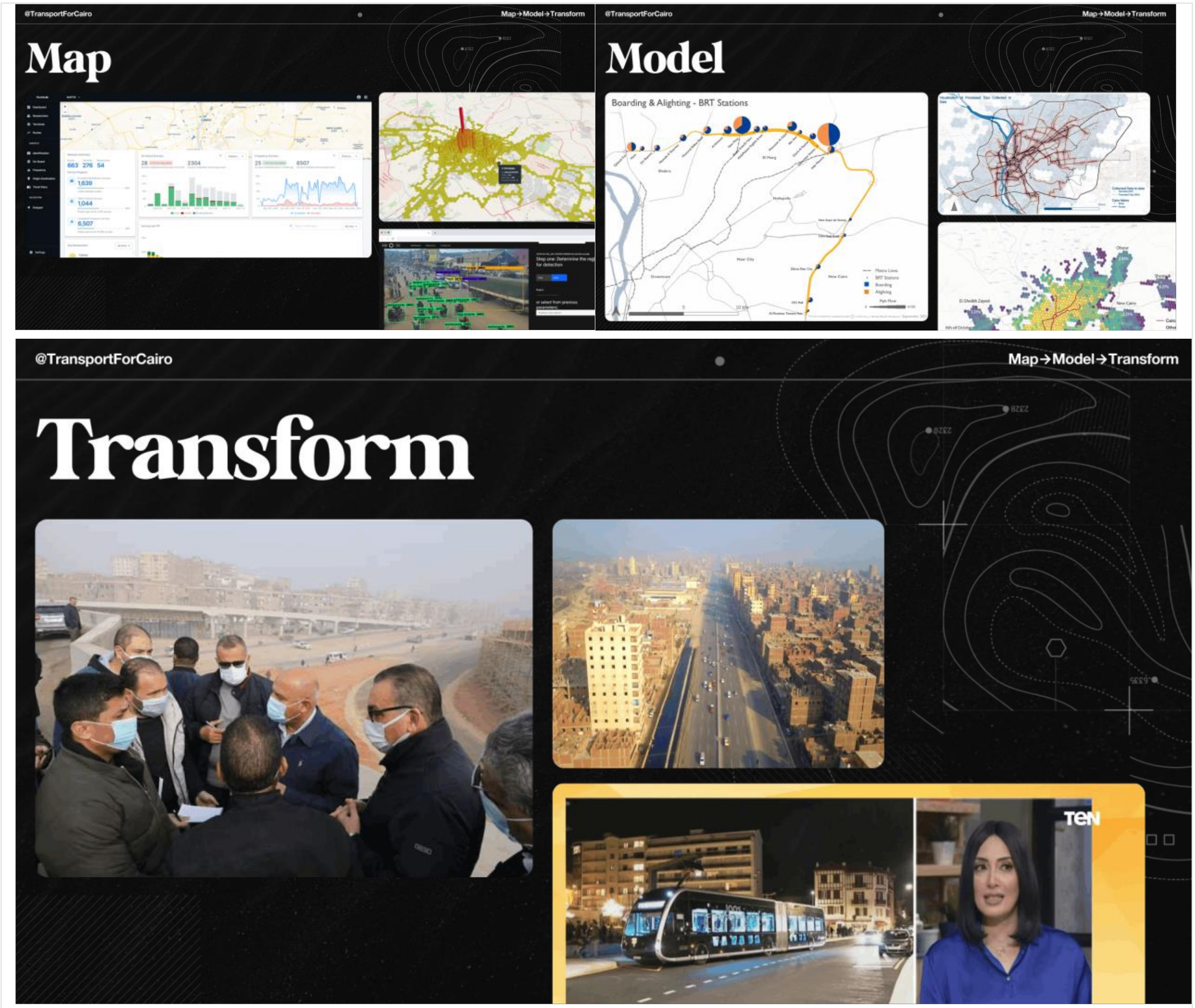

Figure 17: Excerpt from Session 3, Map-Model-Transform Framework

The session then showcased five use cases that can be briefly summarized as follows:

1. **Too many Boda-bodas**: Customizing an object detection model to be able to understand the boda bodas, which were too many to count using human field researchers. The application utilized CCTV video feeds from Kampala's Capital City Authority (KCCA) to automate classified traffic counts
2. **To commute or not to commute**: Using GTFS to model accessibility to jobs using public transport, the use case focuses on the aspect of working within a data scarce environment

3. **Build a BRT**: The application of boarding and alighting data from onboard surveys to determine locations of a planned BRT bus route
4. **At a glance**: The usage of GIS data coupled with graphic design and user experience skills to come up with passenger information systems for Cairo Metro's "Green Line". Those include schematic maps for the metro lines, neighborhood maps for each station, and fare maps to illustrate zone fares
5. **Google it**: Using GTFS to model paratransit and its integration into google maps to enable the "transit" option in Google maps app and allow for journey planning for public transport for the first time in Egypt.

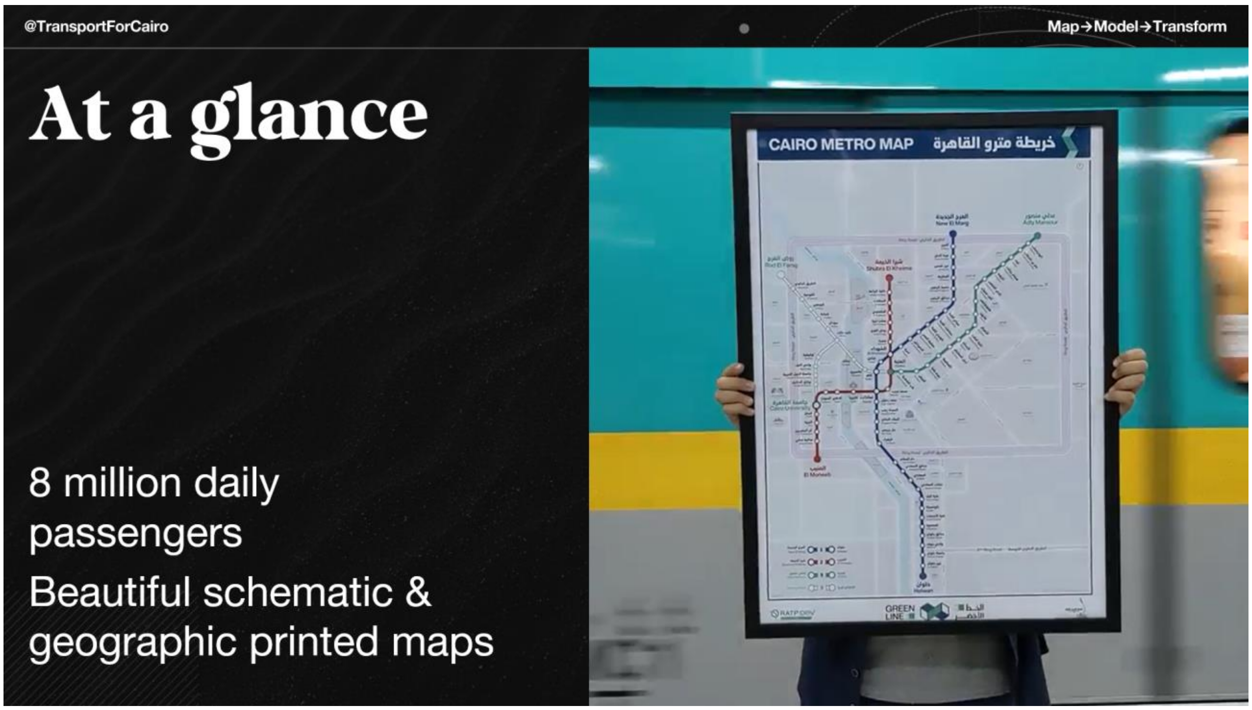

Figure 18: Excerpt from Session 3, Use-case #4 Passenger Information at a Glance

6 Conclusion

6.1.1

6.1 Lessons Learned

GTFS Availability versus GTFS Quality

GTFS as a standard is quite robust in its ability to model and express key elements of a transport system. It is also quite flexible in its core requirements. Meeting the bare minimum requirements of the GTFS specification only enables basic geospatial visualizations. Furthermore, the mere presence of geospatial data is no guarantee of its accuracy; a GTFS feed with duplicate stops and geographic points that are not aligned with the underlying road network will not be as useful as a GTFS with these post-processing steps taken into consideration. With sufficiently high-quality geographic data, we could have conducted two further analytical activities such as:

- **Identifying priority corridors:** whereby spatial itineraries of routes with the highest degree of overlap would represent the potential for introduction of a high-capacity public transport intervention such as a Bus Rapid Transit System or a Light-Rail Tram.
- **Identifying high volume hubs:** similar to priority corridors; analysing the boarding and alighting patterns at each stop - as well as the total number of routes passing through a given stop - could provide opportunities for consolidation into larger "hubs." Further studies and profiling of each stop could clarify opportunities for improving service and passenger safety.

The real power of the specification is when it is combined with accurate temporal data, and multiple agencies to support **multi-modal journey-planning** and accurate **passenger information systems**.

6.1.2 Data Maturity Rating System

While the team set out with the intention of conducting analytical activities on the accrued data to provide actionable insights to the Department of Transport in KMA, the quality of the data along with the client's existing plans required us to pivot to better meet their needs. Perhaps the critical takeaway here is the need for an industry-standard **rating system** based on a **data maturity model** – such as the one developed and presented by TfC in 2.4. This would streamline communications and align expectations on upcoming projects. Additionally, it is prudent to take into account the age of a GTFS feed into this rating system. as TfC noticed that GTFS feeds will tend to "expire" after a number of years without update. This is primarily due to the highly evolving, demand-responsive nature that is characteristic of informal transit.

6.2 Future Work

6.2.1 RouteLab for ongoing Data Collection

RouteLab is a comprehensive, evolving cloud-hosted solution provided by Transport for Cairo to empower the development of data-driven solutions to complex urban mobility problems. To date, cities have leveraged the power and simplicity of RouteLab to map 4 megacities across Africa; including Cairo, Egypt, Accra, Ghana, Kampala, Uganda, and Addis Ababa, Ethiopia. The development of RouteLab is guided by years of data collection expertise within the informal context, as well as constant feedback cycles with cities. Manual data collection using digital tools and field researchers remains to be one of the most effective interventions in building a concrete understanding of this system's operations and patterns.

As part of the DT4A project, TfC agreed to provide continuous access to RouteLab to the KMA for the duration of the year 2022. All training materials were delivered to the KMA team shortly after the workshop, and email support is open and extended until the conclusion of the project should the KMA team decide to conduct their own data collection efforts.

RouteLab is designed for the sole purpose of optimizing the "time-to-GTFS" metric. We measure this metric as the time between the start of a project on RouteLab's Observer, and the conclusion of the data collection effort. We then calculate the time it takes to produce a valid GTFS export from RouteLab's Port, and calculate the cumulative time between field-effort and GTFS provision.

Informal Transit Integration into Ongoing Megaprojects

With the city of Kumasi looking to introduce mass transit systems such as the BRT, it's refreshing to see the staff's efforts into gaining data on their city's paratransit services.

One of the biggest barriers to accounting for paratransit services when implementing mass transit and
6.2.2 infrastructure projects is the lack of visibility. This same lack of visibility is a main reason to the disregard and sometimes mistrust from authorities and decision makers towards those services.

The activity of mapping is the first step towards these services being visible, and thus having insights into how they operate and having the lost jigsaw in the puzzle. However, while this report focuses on mapping and quantitative data, it is also critical to hear out the actual stakeholders (passengers, operators, asset owners, etc.) through the more qualitative surveys and focus group discussion formats.

Introducing change requires a full understanding first of the status quo, and in Kumasi as in Cairo and many other African cities, the status quo has a significant paratransit presence. There will always be the need to strive for data quality and expertise from multiple domains to make this change beneficial for all.

Manual Processing of Raw Boarding Locations Dataset

6.2.3 As mentioned in 4.3.2, the output of the data transformation activity is not fully ready for integration into the Transit Info Map system. The raw points need to be consolidated into a set of locations that cover the main "terminals" or transit hubs where passengers board and alight. While we tried automated clustering methods, we concluded that a manual process is required, based on experience with the local routes.

7 Appendix A: Benchmarking: Three-Pillar Model

To properly assess data quality and comprehensiveness for different cities and in different contexts, TfC developed a "Data Maturity" model to act as a benchmark.

We define the data maturity model as following three pillars of data availability:

1. Geographic Data
2. Temporal Data
3. System Adequacy Data

Table 2: Transit Data Maturity Model

Code	Levels (Sub-categories)	Metadata
Geographic Mapping		
1.1 **BASIC**	Trip GPS-trace	- Route Number / Name - Trip Head-sign
	Trip Stops (i.e. **where** did the bus stop during the trip)	- Stop Location
1.2 **INTRM.**	Trip GPS-trace with basic temporal data	- Route Number / Name - Bus Type & Capacity - Route Description - Trip Fares - Trip Fare Variability 1.0 (YES/NO)
	Trip Stops (i.e. **when** did the bus stop during the trip)	- Stop Name - Stop Type (Planned or Actual or Both) - Time of Trip Start - Trip Duration (between stops) - Dwell Time (at stops) - Time of Trip End
1.3 **ADV**	Multiple Trip GPS-Traces for same route with detailed temporal data	- Route Deviation (Alternative Routes due to congestion, etc.) (YES/No)
Temporal Mapping		
2.1		- Frequency of service (Headway value per route)

Code	Levels (Sub-categories)	Metadata
BASIC		- General Operating Time
2.2 INTRM.		- Operating Time Variations (Seasonal Schedules, …etc.) - Availability of Passenger Information System 2.0 (in Stops)
2.3 ADV		- Frequency Variation per stop per Vehicle (x Vehicles "same route number or name" / stop / hour)
System Adequacy		
3.1 BASIC	End of Trip Survey	- Availability of Safety Equipment and measures (Door closing, Standing Bar … etc.) - Availability of Passenger Information System 1.0 (in Vehicles) - Emergency Preparedness (Vehicle) - Wheelchair Accessibility
3.2 INTRM.		- Crowding 1.0 (General Score based on trip crowding level) - Use of Renewable or Clean Energy - Comfort Measures (AC availability, comfortable seating … etc.) - Stop Amenities (Availability and Quality) - Emergency Preparedness (in Stops)

Code	Levels (Sub-categories)	Metadata
3.3 **ADV**	General	- Crowding 2.0 (Per Stop to measure passenger flow in and out of the vehicle) - Safety and Walkability of Stop's catchment area (Built Environment Infrastructure Data) - Trip Fare Variability (Value by distance) - Energy Efficiency - Noise Pollution Level - GHG And Pollutant Emissions

8 Appendix B: Existing Data Assessment

This section provides an assessment of the dataset pertaining to Kumasi, Ghana. We assess the quality of the data and the availability of complementary data, as well as present potential applications for it.

Provided Data Overview

GTFS Feed

8.1.1 The data provided consists of two sets of files: A GTFS feed, and a set of excel sheets detailing various route legs.

The GTFS feed breaks down into the following tables:

```
Agency: 1

1. Main tables: (*)
agency              : 1 entries
calendar            : 1 entries
stops               : 704 entries
routes              : 668 entries
trips               : 668 entries
stop_times          : 2,519 entries

2. Additional tables: (#)
calendar_dates      : 0 entries
fare_attributes     : 0 entries
fare_rules          : 0 entries
shapes              : 114,512 entries
frequencies         : 0 entries
transfers           : 0 entries
feed_info           : 0 entries

----
*: required part of GTFS spec, needed to make valid GTFS
#: part of GTFS spec but not compulsory
```

Excel Sheets

The excel sheets consist of a variety of Origin-Destination pairs of points. The exact path of each raw trip is specified in the accompanying “Data” sheet, interestingly, a Google routing link is also calculated between each OD pair. The presence of the Google links requires some further clarification, as it raises question of whether field research was conducted, or if OD pair links were simply sampled.

Third-Party Data Assessment

Administrative Boundaries

Administrative boundaries were available at three levels of resolution courtesy of the United Nations Office for the Coordination of Humanitarian Affairs[4] (UN OCHA). The highest resolution layer
8.1.2 attributed 43 geographic areas to the "Ashanti" region, in which Kumasi resides at the heart of.

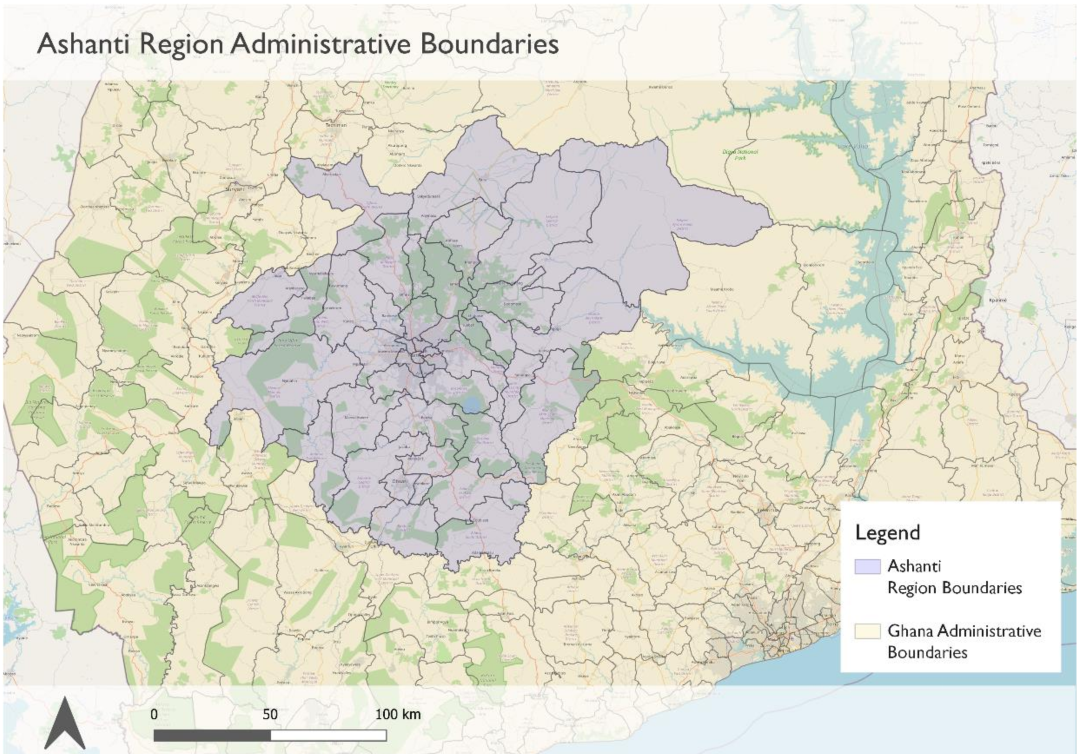

Figure 19: Ashanti Region Administrative Boundaries subset of the Ghana Administrative Boundaries

Population Density

8.1.3 Population data acquired from "WorldPop"[5] is at a resolution of 100m per grid cell. The data is obtained using a Random Forest regressor model, used to disaggregate sensed and ancillary data.

Validity

Beginning with the GTFS feed agencies and feed info, we can already identify a few issues:

1. There is no feed metadata,

[4] "Ghana - Subnational Administrative Boundaries", https://data.humdata.org/dataset/cod-ab-gha

[5] "The spatial distribution of population in 2020, Ghana", DOI : 10.5258/SOTON/WP00645, WorldPop (www.worldpop.org - School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University (2018). Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation (OPP1134076).

2. The agency name is either codified or is otherwise unhelpful

One interesting point to note is that the number of routes and trips are identical. Upon further inspection, the number of routes is artificially inflated as there are routes for each leg. The GTFS spec leverages a "direction" flag to signify outgoing and return journeys, it is not used here; instead, separate routes are created for each leg of the route. This means the dataset only consists of **334 routes**.

Running this feed through a validator produces the following results:

	Error Count
High Severity	1
Medium Severity	5,141
Low Severity	0

The detailed overview provides a bit more information on the different kinds of errors present. These include the following:

Severity	Error Type	Error Count
High Severity	Travel time negative	1
Medium Severity	No service on a date	4,068
Medium Severity	Route short name too long	668
Medium Severity	Duplicate stop	216
Medium Severity	Stop unused	186
Medium Severity	Travel too fast	2
Medium Severity	URL format	1

High severity errors are classified as such if the error is likely to break routing results. For example, if stop times are in an incorrect sequence, or the time between stops denotes unrealistically high-speed travel this would fail to route. In the case of the provided GTFS feed the link between two stops shows a negative travel time. High severity errors such as the one present would not be usable in trip planning applications.

Medium severity errors tend to break the display of data, which leads to a poorer user experience; however, it will not affect routing results. The clear anomaly here is the 4,068 trip entries which are missing a corresponding service_id to denote their operating window. This indicates a lack of temporal data across the dataset.

Of the 704 stops present in the GTFS feed, 216 of them are identified as duplicates – with an additional 186 (26.4% of stops) identified as unused by any route or trip path.

A quick visual inspection gives us a better idea of the stops' distribution. We see a pattern – similar to Cairo as a matter of fact – with an exceptionally busy city center and a multitude of stops around the main Bypass which encircles the city.

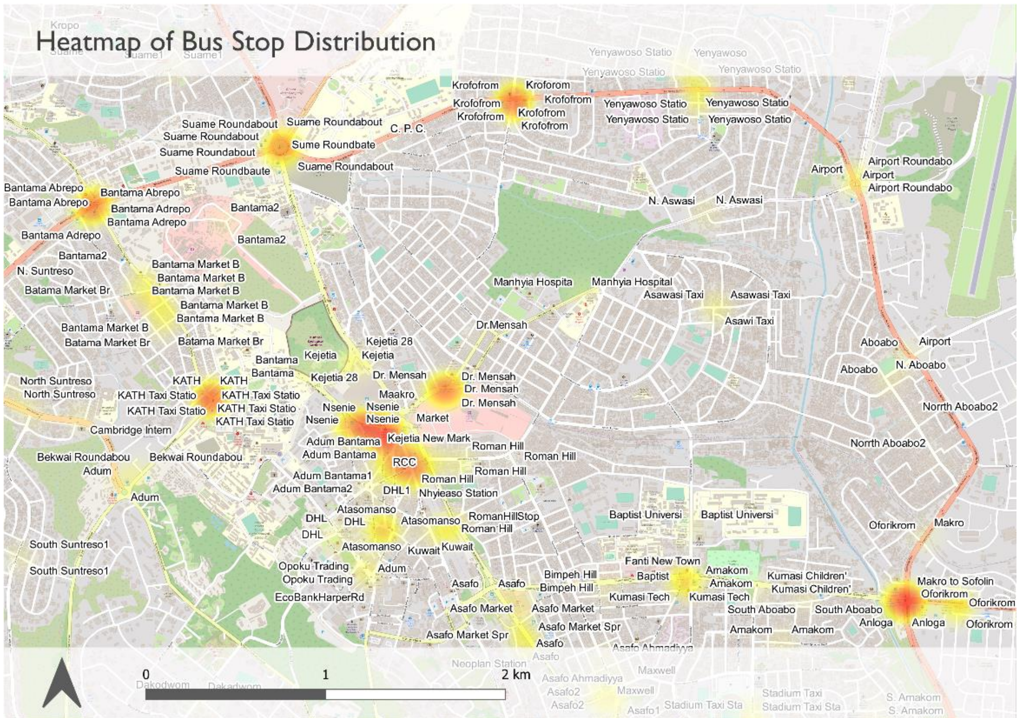

Figure 20: Stops Heatmap, Central & P.V. Obeng Bypass

Zooming in to these 'hotspots' we can quickly identify examples of duplicate stops as shown in Figures 3 & 4.

Figure 21: Duplicate clusters of stops

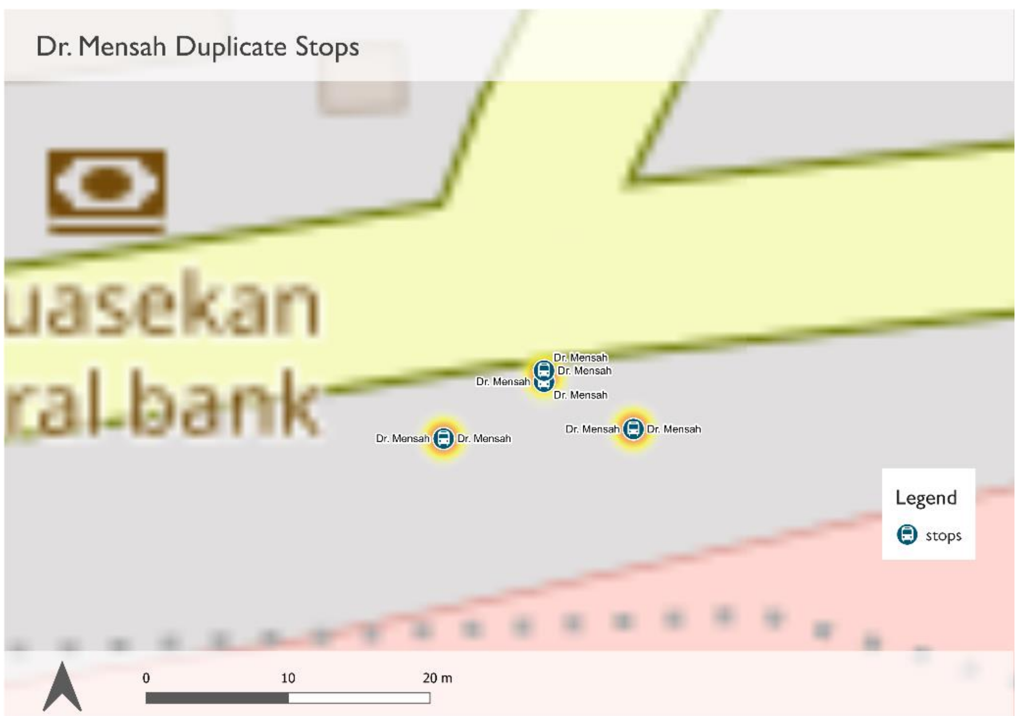

Figure 22: Dr. Mensah stops overlap geographically and are not aligned to the roads layer

Not only are some stops duplicated, but it seems that some stops are not aligned to the nearest road boundary. Within a 20m radius, we see stops like Dr. Mensah which overlap one another; sometimes completely (notice a single marker will display a stop name multiple times due to the underlying duplicate).

Currency of the Data

Although the GTFS feed lacks the metadata detailing currency, the raw data supplied is named according to the dates at which each set of routes was collected. The data is accurate to Kumasi for the month of March of 2020.

8.1.4 WorldPop's population dataset was created in 2018 and has been updated since to provide population distribution estimates for the entirety of Ghana for the year 2020.

Data Quality

8.1.5 Geographic Coverage

Within the Ashanti region, the routes provided in 2.1.1 extend to cover approximately 15 areas.

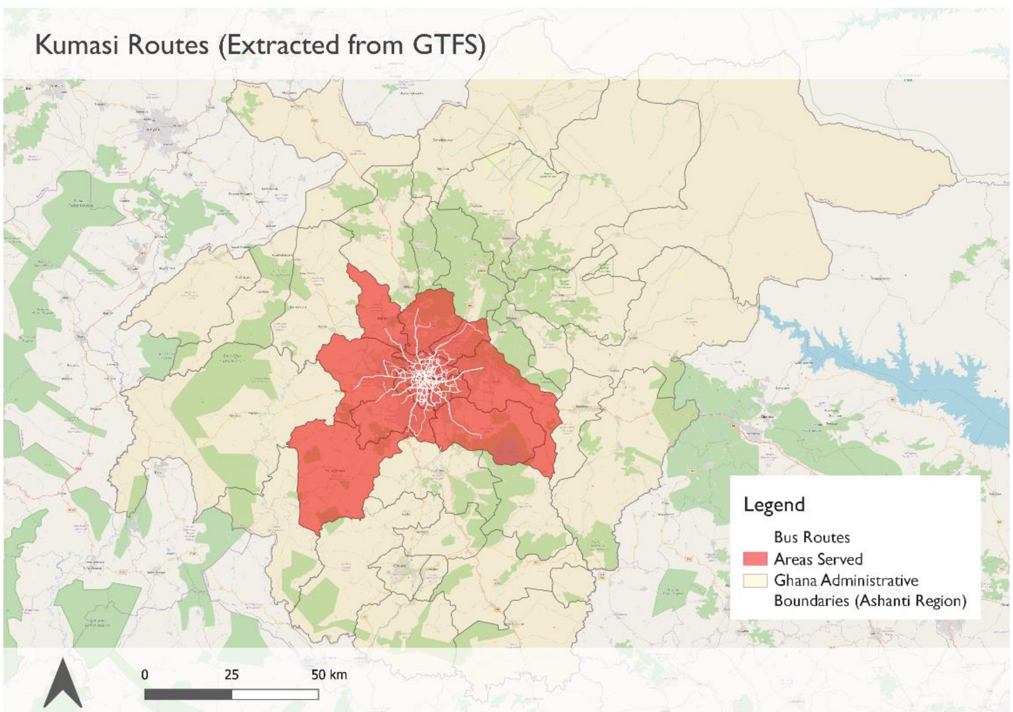

Figure 23: Districts of the Greater Kumasi Area

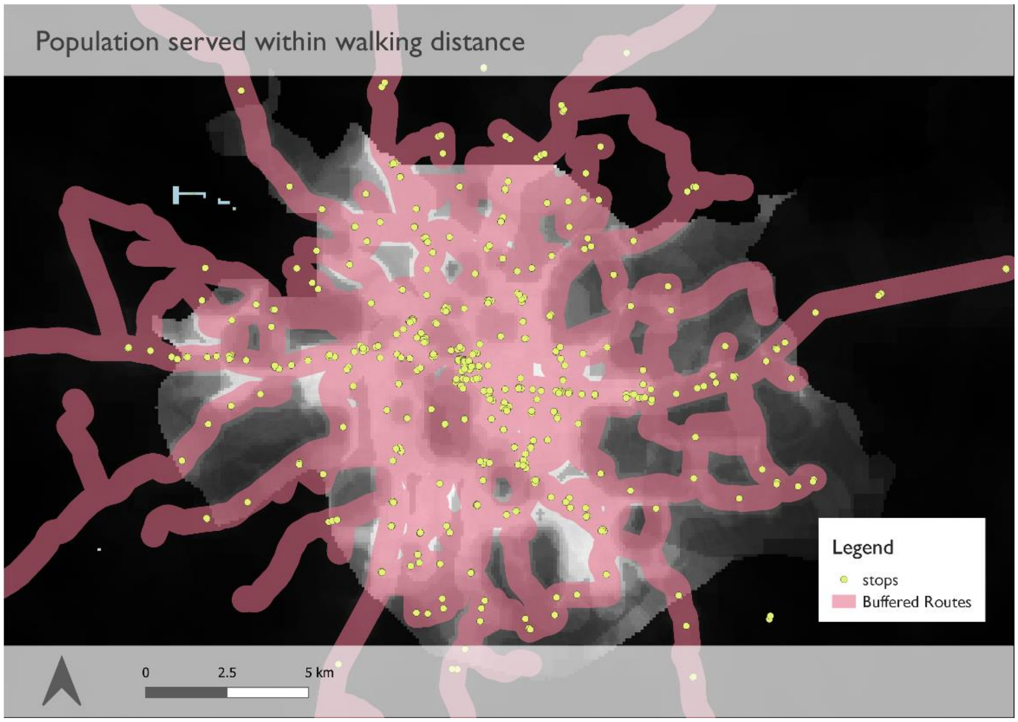

Figure 24: Population served within 400m radius around trip paths

A quick overlay analysis shows that in the Kumasi region (with a population of approx. 3.348 million inhabitants) approx. 2.89 million are within a 400m radius from the nearest trip.

Temporal Coverage

The stop_times table from the GTFS feed includes only basic temporal data about each trip. Namely, the duration of the trip – using arrival and departure times for the first and last stop of any given trip. However, all mid-stops are void of any temporal data.

Furthermore, mid-stop times cannot be estimated due to the missing frequencies table which would have provided a headway estimate at minimum.

This complete lack of the temporal dimension makes it difficult to extract any information of the supply-demand characteristics of each trip.

8.1.6 Metadata

Unfortunately fare data also happens to be missing from the provided data.

Summary

As per our Data Maturity Model, the Kumasi dataset covers the following items:

- Geographic Mapping Data
  - ✓ [BASIC] **Route Number/Name**
  - ✓ [BASIC] **Stop Location**
  - ✓ [INTRM] **Stop Name**

✓ [INTRM] **Trip Duration**

A lack of Temporal and System Adequacy data is observed. As such, a full-scale data collection effort is highly recommended.

Using the FR Planning Tool, we estimate a total required FR Capacity of 1,975 person-days to map 668 Unique Routes (both onboard & frequency surveys). This figure is derived from the number of Route Legs and the average FR performance (for each survey) based on the historical experiences of TfC in the field.

To map 668 routes (i.e., 1,336 route legs), we would require:

- 432 person-days of Onboard Surveys
- 1,543 person-days of Frequency Surveys

If we assume a field research capacity of **20 researchers**, we can collect the necessary 2,672 **onboard** assignment instances (based on a redundancy of 2 surveys per route leg) in **23 days**; and the same number of **frequency** assignment instances in **78 days**.

Before conducting a mapping exercise however, a preliminary identification activity must be conducted to validate the existence of 668 unique routes. Due to the nature of the provided data, we believe there are only 334 unique routes. If that is the case, the estimates would drop to **13 and 40 days** for onboard and frequency collection respectively.