EPFL HUMAN RESEARCH ETHICS COMMITTEE

RESEARCH PROTOCOL AND ETHICAL ISSUES - SECTION C2

This document needs to be uploaded in section C2 of the on-line HREC application form

Name applicant	Franziska Zollner
Title project	Impact of the Tanzanian Standard Gauge Rail on Economic Development

Please describe your research protocol, incl. study design / research methodology (e.g. observational, survey research, experimental). Give details of any samples or measurements to be taken.
This project uses observational, non-experimental data to study the impact of the opening of Tanzania's Standard Gauge Railway (SGR) on mobility patterns and economic activity. The research design exploits temporal and spatial variation generated by the opening of the railway to compare mobility flows and economic indicators before and after the start of operations, as well as across locations differentially affected by the new transport connection. No intervention or direct interaction with human subjects is involved. The primary data source is mobile phone location data purchased from the data provider Quadrant. The data cover users in Tanzania over a period of several months surrounding the opening of the SGR, potentially before, during and after the opening of the trainline (the exact time window will be determined based on data availability). The unit of observation in the raw data is a location "ping" generated by a mobile device when an application with location permissions records a GPS event. The dataset includes the following variables: a device-level identifier (mobile advertising ID), identifier type (iOS IDFA or Android ADID), timestamp, latitude and longitude coordinates, GPS accuracy, device operating system and OS version, user agent string, country code, IP address at the time of the event, application and publisher identifiers, indicators for foreground/background app activity, geohash, and indicators that user consent was obtained for data collection. The data are provided without names, phone numbers, or other directly identifying personal information. In this project, I do not analyze individual-level behavior. All analyses are conducted on aggregated and anonymized measures constructed from the raw GPS data. No attempt is made to identify, profile, or track specific individuals or households. From the GPS data, I intend to construct aggregate mobility measures at the level of geographic areas (e.g., grid cells, transport corridors, or administrative units) and time periods. These measures include, but are not limited to: average travel frequency, average distance travelled, and origin–destination flows between locations. To study economic outcomes, the mobility measures are combined with publicly available remote-sensing data, such as nighttime light intensity and built-up area indicators, which capture changes in economic activity and capital accumulation over time. No linkage to administrative records or personal datasets is planned.

Please describe the ethical issues linked to your research project and how they will be addressed.
This project involves the processing of personal data within the meaning of the Swiss Federal Act on Data Protection (FADP). The mobile phone location data include device-level identifiers and precise GPS coordinates. The main ethical concern is the risk of identification and potential misuse of fine-grained mobility data. This risk is addressed through several safeguards. First, the research does not analyze individual-level behavior. All outputs are strictly aggregated across users, space, and time, and no results are reported for small geographic areas or small groups that could enable inference about individuals. The analysis focuses exclusively on broad mobility patterns and spatial connectivity at an aggregate level. Second, consent is obtained upstream by the data provider. The mobile location data are collected by applications that obtain opt-in consent directly from users, typically at first launch. The researcher does not collect data directly from individuals and has no contact with data subjects. Third, all data are handled in accordance with data security and access control standards. The data is planned to be stored on servers physically located at the EPFL campus on a NAS with access only for project participants. The data will be deleted 5 years after the thesis of the students working on this project are submitted and approved. The data are used exclusively for academic research purposes and are not shared with third parties. The data will remain private and confidential and will not be shared anywhere. Finally, the risk to data subjects is minimal. The research does not involve interventions, deception, or interaction with individuals, and the findings are reported only in aggregate form. Given the safeguards in place, the likelihood of harm to individuals whose data are included in the dataset is considered very low.