Establishing a Secure Data Center with Remote Access

Jeffrey Gonder[1], Evan Burton[1] and Elaine Murakami[2]

1 – National Renewable Energy Laboratory (NREL), Center for Transportation Technologies and Systems (CTTS); 2 – U.S. Department of Transportation (DOT), Federal Highway Administration (FHWA)
Jeff.Gonder@nrel.gov; Evan.Burton@nrel.gov; Elaine.Murakami@fhwa.dot.gov

Abstract

Access to existing travel data is critical for many analysis efforts that lack the time or resources to support detailed data collection. High-resolution data sets provide particular value, but also present a challenge for preserving the anonymity of the original survey participants. To address the dilemma of providing data access while preserving privacy, the National Renewable Energy Laboratory and the U.S. Department of Transportation have launched the Transportation Secure Data Center (TSDC). TSDC data sets include those from regional travel surveys and studies that increasingly use global positioning system devices. Data provided by different collecting agencies vary with respect to formatting, elements included, and level of processing conducted in support of the original purpose. The TSDC relies on a number of geospatial and other analysis tools to ensure data quality and to generate useful information outputs. TSDC users can access the processed data in two different ways. The first is by downloading summary results and second-by-second vehicle speed profiles (with latitude/longitude information removed) from a publicly accessible website. The second method involves applying for a remote connection account to a controlled-access environment where spatial analysis can be conducted, but raw data cannot be removed. The TSDC website is http://www.nrel.gov/vehiclesandfuels/secure_transportation_data.html.

Background

Travel behavior surveys have been conducted for many decades and have been a critical component for long-range transportation planning. Historically, these surveys used self-reported paper diaries with telephone retrieval, but global positioning system (GPS) technology has been used starting in 1996 [1]. Since 1995, lowered costs of GPS equipment and the incorporation of GPS into smartphones have made a GPS component nearly standard practice for travel behavior surveys. Even if a specific address is not provided on a survey record, a GPS data logger available for less than $30 can provide sufficient spatial resolution to identify a specific house or store location. The resulting data enhancements, such as improved trip reporting rates and increased temporal and spatial resolution associated with GPS surveys, have made the collected information even more valuable for traditional uses and have also created an opportunity for expanded uses of the data beyond its original purpose [2–5]. However, researchers collecting this kind of information need to assure respondents that anyone accessing the survey results will not disclose information about a specific person or otherwise use it nefariously.

The costs to collect this type of data can be significant. For value pricing studies using in-vehicle equipment to track hundreds of vehicles over several months (such as those completed recently in Seattle, Atlanta, and Portland, Oregon), the data costs can total *one to two million dollars for a single study*. There have also been a number of recent instances where a metropolitan planning organization (MPO) augmented its regular paper survey with a GPS component (such as in St. Louis, Los Angeles, Kansas City, Chicago, Washington, D.C., San Antonio, Austin and Houston, to name a few). In these cases, the GPS add-on covered from a few hundred to one or two thousand vehicles, driving anywhere from one day to a couple of weeks. Such *add-on GPS survey components still cost several hundred thousand dollars each* to collect. In spite of the significant resources spent to collect this high-resolution data, concerns over potential misuse have made most organizations reluctant to share it. As a result, although valuable and costly to collect, the data are being underutilized and in some cases becoming "mothballed" or scheduled for destruction.

Traditional household travel behavior survey data have been collected and stored in the Metropolitan Travel Survey Archive (MTSA) at the University of Minnesota [6]. The MTSA has received intermittent funding from the U.S. Department of Transportation (DOT) through the Federal Highway Administration (FHWA) and start-up funding

from the Bureau of Transportation Statistics (BTS). However, transportation agencies that had incorporated GPS into their travel behavior studies were reluctant to provide a copy of the GPS data, even if only for archival purposes and not for data access.

After considering how best to resolve the conflict between data utilization and confidentiality protection, a 2007 report by the National Research Council recommended a data enclave solution [7]. The National Renewable Energy Laboratory (NREL) has implemented a similar approach with the Hydrogen Secure Data Center (HSDC) since 2003. The HSDC provides access to data on prototype fuel cell vehicle and hydrogen fueling station deployments, while protecting individual details on each automaker's vehicles [8]. NREL has also worked extensively with GPS travel data and received numerous data requests following publication and presentation of associated research [2].

Recognizing the value for their own analysis efforts and for the greater research community, NREL and FHWA began partnering in late 2009 to develop a data center that allows access to highly detailed records of travel in time and space in a way that maintains respondent anonymity. NREL leveraged existing security infrastructure and staff from the HSDC, and contributed internal funding to purchase equipment for the new data center. The FHWA Offices of Planning and Operations jointly contributed funding to support startup and initial operation of the Transportation Secure Data Center (TSDC) that has now been established. The remainder of this paper provides further details on the approach, structure, and contents of the TSDC.

Approach

In addition to building off of NREL's experience with data centers such as the HSDC and the Alternative Fuels and Advanced Vehicles Data Center (AFDC) [9], FHWA and NREL considered other secure data centers while formulating the approach for the TSDC. For instance, the Census Bureau's Research Data Center (RDC) program has a long-established system of providing researcher access to highly confidential data [10]. One way that the RDC program manages data security is by requiring users to travel to specific locations to access the data. Although the Census Bureau has continued to add new RDC sites, this travel requirement presents an inconvenience to many researchers. The NORC Data Enclave provides one example of a repository that permits restricted remote access for researchers [11]. The NORC Data Enclave stores social science micro-data records, such as from the Annie E. Casey Foundation's Making Connections Survey on topics such as economic hardship in families, and from a National Science Foundation Survey of Earned Doctorates.

While designing the TSDC, it was decided that having secure remote access to a single data center would give the greatest benefit to both data providers and users. On the provider side, this relieves the burden on each MPO or transportation agency of having to store and protect data, respond to data sharing requests, and/or set up their own secure data center (some agencies have in the past provided their data to others on a case-by-case basis). Users also benefit from not needing to travel, from finding data in a single location, and from working with data they would not otherwise be able to access. As described later, accomplishing this vision required implementing technical controls to prevent removal of data through the remote connection and providing appropriate tools on the remote site for use in conducting analyses.

An important step in the TSDC development process was establishing an advisory group to help provide oversight, to give technical input, and to represent the interests of various stakeholders. This group includes data providers and users who work in industry, academia, and government. Input from the advisory group helped drive the TSDC development philosophy to first of all implement security measures for protecting data and minimizing or eliminating the potential for privacy violation. The users represented on the advisory committee helped make sure these protections were implemented while still allowing researchers to access critical data elements in a user-friendly manner.

To help strike the balance between protection and usability, NREL organized the TSDC into three distinct sections: (1) a secure enclave for raw data, (2) a downloadable area for cleansed data and (3) a controlled-access area for detailed data. Spatial details have been removed from the cleansed data in the downloadable area, preventing identification of individual participants. This approach offers strong data protection while satisfying the needs of many users who simply require aggregated driving information or individual vehicle speed profiles. The downloadable area allows these users quick and easy access to the data they need, without requiring them to go through the more involved approval and connection procedure for the controlled-access area. Data security is also

enhanced by not providing access to more data details than these users actually require. Additional details are provided later in this paper on the three TSDC sections, and how each furthers the dual goals of data protection and usability.

Example Datasets and Uses

Some of the TSDC's datasets were collected as add-on samples to traditional diary-based regional travel surveys for assessing issues such as trip underreporting. These GPS samples typically include second-by-second data for several hundred vehicles recorded over one or more days—up to a week or two at most. The Texas Department of Transportation provided such travel datasets for Houston, Galveston, San Antonio, and Austin (variously collected from 2005 to 2009).

The Southern California Association of Governments (SCAG) provided a similar travel dataset for a study conducted in the Los Angeles area from 2001 and 2002. SCAG analyzed raw GPS points to approximate participant home, work and trip start/end locations by using the center of the U.S. Census tract in which these points fell (the Census Transportation Planning Package uses a similar approach to aggregate data). SCAG then provided a dataset to the TSDC that includes tract-level summaries and corresponding data files as well as full GPS travel profiles.

The Puget Sound Regional Council (PSRC) in Washington provided another large dataset from a FHWA-sponsored research project on pricing. The Traffic Choices Study collected data from 2004 through 2006 to evaluate travel behavior changes in response to time- and location-variable road tolling. The particular value of this dataset is that it contains GPS data from over 400 vehicles over a period of 18 months, including baseline data collected prior to the seven-month experimental tolling period.

A research group from the University of Texas at Austin was one of the first to publish about an analysis using data accessed from the TSDC [12]. Khan and Kockelman used data on daily travel variability from the Traffic Choices Study to estimate how well a battery electric vehicle with a 100-mile range could serve high shares of American households with minimal adjustments (e.g., borrowing a car every 100 days). They found that 50% of single-vehicle households and 80% of multiple vehicle households would be satisfied under these conditions.

The battery electric vehicle analysis is just one example of the many potential applications (both traditional and non-traditional) that can benefit from using this type of detailed travel data. Table 1 provides a partial list of users and use cases that can take advantage of past and future GPS surveys as the TSDC effort continues. Figure 1 shows an example geographic information system (GIS) analysis made possible by TSDC data.

Table 1. Example Users and Use Cases for GPS Travel Data

Example Data Uses	Example Data Stakeholders
• Transit planning • Travel demand modeling • Congestion mitigation • Vehicle energy and power requirement analysis • Air quality and emissions modeling • Disaster/evacuation planning • Spatial/GIS analysis (for alternative fuel station planning, accessibility assessment, etc.) • Climate change impact studies	• DOT (FHWA, Federal Transit Administration, etc.) • DOE and National Laboratories • U.S. Environmental Protection Agency (EPA) • Air Resources Board • Cities, MPOs & State DOTs • Universities • Auto Manufacturers • Department of Homeland Security

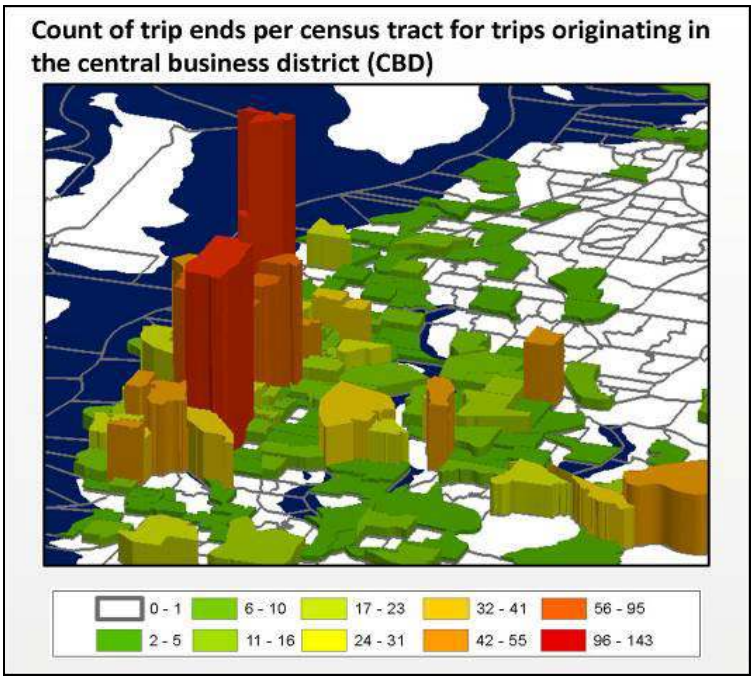

Figure 1. 3D Rendering of an example accessibility analysis using TSDC data

TSDC Section 1: Secure Enclave for Raw Data

As mentioned in the Approach discussion, the TSDC is organized in three sections. The first section is a secure area where raw datasets from providers are archived and backed up. This area incorporates the significant data protection measures established by the HSDC:

- Located in a building with an on-site security force and accessible only by badge access
- Server room requires separate badge and personal identification number (PIN) access for authorized users only, and possesses its own motion detector and alarm system
- Equipment uses an isolated network, with no connection to the internet or NREL's site-wide network
- Data mirrored on multiple terabyte storage array
- Weekly tape back-ups stored in server room fire-proof safe, and monthly back-ups stored in another building

Processing performed on workstations in this area checks data for quality, corrects errors, and prepares versions of each dataset for posting in the other two sections of the TSDC. To prepare externally accessible versions, the first step is to locate and remove any explicitly identifying participant information (such as respondent names and addresses that may have been included in the dataset sent to NREL). NREL takes several additional privacy protection measures to prepare data for posting in the TSDC's downloadable area for cleansed data. These steps include removing latitude-longitude details from the data, as well as other information (such as vehicle model) that could potentially be used in combination with other details to identify a participant.

Data quality control steps include identifying and correcting GPS measurement errors, which can result from equipment cold-start (where recording does not begin until after a vehicle/person has already started moving), signal dropout, and measurement drift. If uncorrected, these errors can lead to unreasonable calculations of driving distances, speeds and accelerations in second-by-second GPS data. The data are also scrutinized to identify outlying values and recording errors that may or may not have been flagged in the original study. For instance, in one study about two percent of trips were inadvertently assigned duplicate identification numbers, resulting in huge jumps in the GPS point recordings (actually made by two different vehicles driving across town from one another). In another study 5%–10% of the GPS data collection devices were inadvertently set to record speed data in kilometers per hour rather than miles per hour. When the data was interpreted in the expected unit of miles per hour some of the vehicle speeds appeared high. The error was identified and corrected by making point-to-point distance calculations to compare against the speed readings.

TSDC Section 2: Downloadable Area for Cleansed Data

Figure 2 shows a screen shot of the TSDC website, from which users can download the cleansed data. As described above, the cleansed data includes aggregated statistics, trip driving distances and second-by-second vehicle speed profiles. The cleansed data does not include detailed location and other information that could be used to violate the anonymity of an individual participant. In addition, users are required to register and accept a point-and-click legal disclaimer and use agreement (also pictured in Figure 2) stating that he/she will not attempt to identify individual or personal information from the data.

Figure 2. TSDC website and downloadable cleansed data registration form

After completing the simple registration form and logging in, users can view a more detailed summary description of the studies from which data has been made available. Several files are also readily available to download for each study. These typically consist of:

- The final report – Providing extensive details on the original study and the context of the collected data
- Multiple data files – Typically in comma-separated variable (.csv) format, and generally including:
    - Detailed point-by-point travel data for each trip by each vehicle/participant over the entire study period, both in "raw" form and in a processed form with known errors corrected
    - Demographic data on participants (age, income, number of vehicles in household, etc.) and their vehicles (year, make, EPA functional class, and fuel economy)
- A data dictionary – Providing a general overview of the study and data, any issues of which to be aware, definitions of each variable in the downloadable data files, and summary statistics (on driving distances, speeds, etc.).

Though the data elements removed for privacy protection cannot be accessed from this area, NREL has tried to add other useful details into the data to accommodate common use cases. For instance, while vehicle model information has been left out of the data as an additional privacy precaution, information on the EPA vehicle classification (car, truck, etc.) and fuel economy rating has been added to help support efforts such as air quality analyses and emissions modeling. Other research efforts supported by the type of data available for download include analyzing real-world

fuel economy and powertrain demands on advanced technology vehicles and carbon dioxide emissions impacts of traffic congestion [2–5].

NREL, FHWA and the TSDC advisory group also considered processing options to generate other types of cleansed data. One considered approach was to code trip ends to the centroid of a geographic region such as a traffic analysis zone (TAZ). This would make some spatial data available for download, simply with obfuscated trip ends. However, this option was ultimately turned aside for failing to strike the desired balance between privacy protection and usability. Even partially obfuscated spatial data could present an added privacy concern when placed alongside other details such as driving speed/distance, household/vehicle demographics and trip purpose. Usability of the obfuscated spatial data would also be limited by the inability to link actual trip ends back to land use information from sources such as parcel data and Google Maps. It was instead decided to only provide spatial data through the controlled-access area described below—relying on rigorous user screening, technical controls and a legal agreement for privacy protection, and providing numerous GIS tools and reference data to support user analyses.

TSDC Section 3: Controlled-Access Area for Detailed Data

To help flesh out the details for developing the controlled-access area, NREL, FHWA and the TSDC advisory group again considered related best practice examples. The process used by the National Household Travel Survey (NHTS) for researchers to request the "NHTS DOT file" presented the most similar situation to the TSDC goals—permitting restricted access to enhanced geographic and demographic details (that *could* be abused to identify a participant) in order to permit legitimate research/analyses that could not otherwise be accomplished without the enhanced data. The NHTS procedure requires users to describe the analysis they wish to conduct and sign a confidentiality agreement before receiving the detailed data. The TSDC adopted these as well as additional privacy protection steps as summarized below:

- Applicants must complete an analysis description form describing:
    - The analysis to be conducted
    - Other data sources considered to support the analysis and why they are insufficient
    - The output results the user anticipates wanting to extract from the secure environment
- Applicants must sign a data use and disclaimer agreement including:
    - Confidential data protection legal language
    - An explicit pledge not to attempt identifying individual participants
    - An additional signature required from his/her University Advisor or Line Manager
- Users must also complete a Condition of Use for Cyber Resources form before they may establish a connection account to the NREL virtual machines hosting the controlled-access data
- After users complete the paperwork the advisory group reviews the application and provides an access recommendation; DOT and NREL then make the final approval or denial decision
- Once approved, users may only interface with the data through the controlled-access environment:
    - The environment prohibits data transfer (clipboard sharing, local drive access and internet connection are all disabled)
    - NREL reviews any externally-developed user code or base files before loading it into the environment for use, and similarly audits aggregated results a user wishes to extract before providing them to the user

NREL has again tried to balance operating within the constraints of these security measures with inclusion of features and functions to support usability of the controlled-access area for researchers. These efforts have included providing a variety of software tools to support a range of user preferences and abilities. The available software includes free and open source (FOSS) tools for database query, programming and analysis, such as PostgreSQL/PostGIS/QGIS, uDig, GRASS, Python(x,y) and R. The database-integrated FOSS tools help provide computational efficiency for large GIS analyses (important when a dataset contains millions of GPS data points), but require some programming expertise. Users are therefore also given access to the commercial ArcGIS software from ESRI, which provides a user-friendly integrated system for working with GIS data through a graphical user interface.

To further support an assortment of spatial analyses, NREL has included a variety of additional GIS reference information in the controlled-access area. These data include U.S. Census Topologically Integrated Geographic Encoding and Referencing (TIGER) system files showing water bodies, roads, landmarks and political boundaries

(such as county, census tract and block group). To accompany the Traffic Choices Study data, the TSDC also includes archived road speed and UrbanSIM data for the Puget Sound Region (grid-based information on demographic, economic and land use characteristics). As mentioned earlier, users can ask NREL to load additional tools or reference files into the controlled-access environment in order to support a particular analysis. Figures 3 and 4 illustrate a few examples of the available tools and reference information.

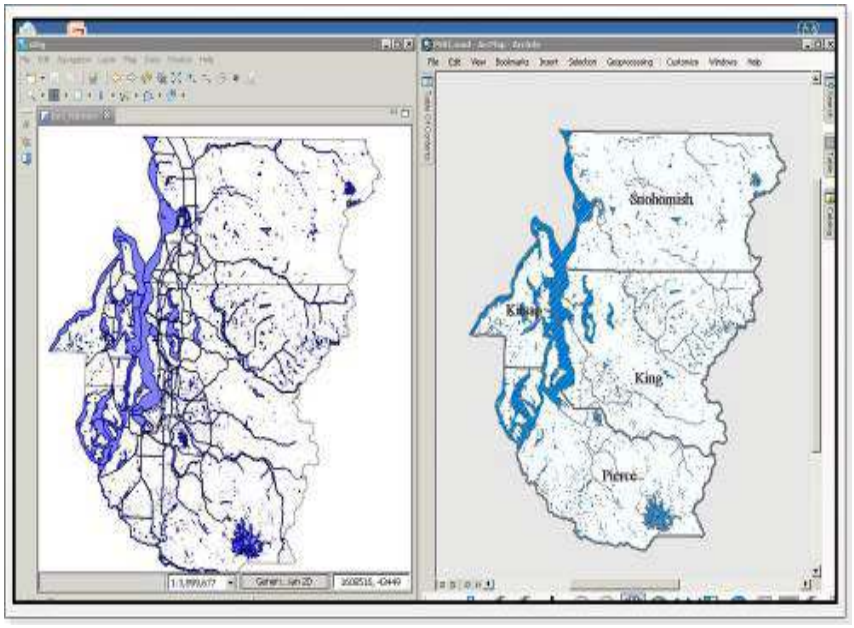

Figure 3. uDig and ArcGIS desktop applications, with reference information on water bodies, roads and county boundaries

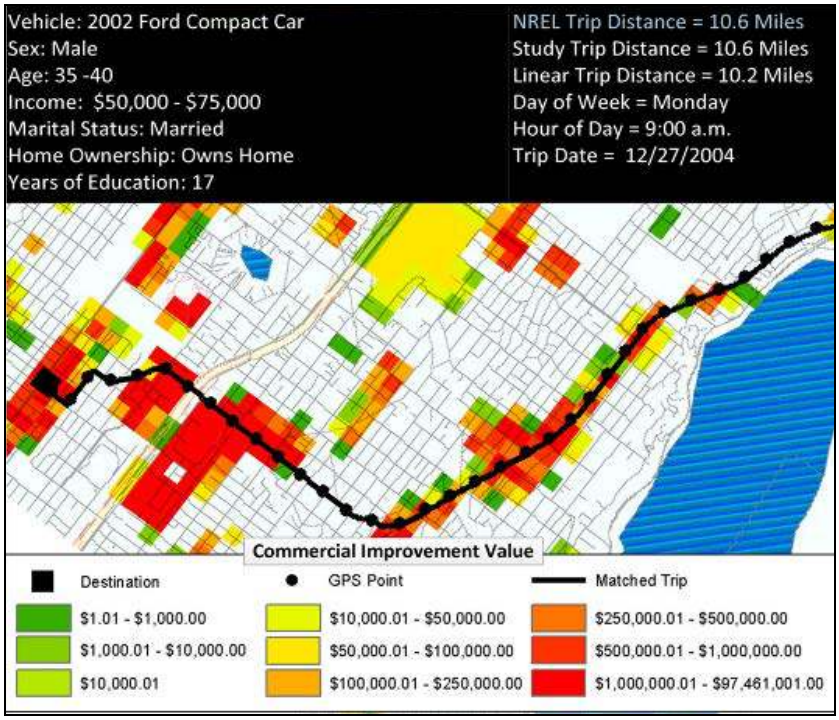

Figure 4. Example of demographic, processed trip and regional data available in the TSDC

Additional efforts to improve the usability of the controlled-access environment include creation of a detailed user manual (navigable by xml within the environment as shown in Figure 5) developed in conjunction with pilot testing of the processes to apply for access, grant approval, remotely connect and conduct analysis. As summarized in a Federal Committee on Statistical Methodology research conference publication, such usability testing can help find problems and identify solutions in an initial system prototype before it begins serving a larger number of end users [13]. Testing conducted by external researchers as well as NREL employees remotely connecting to the environment has helped ensure that:

- Users can connect to the controlled-access site without interruption (users were initially asked to provide the IP address from which they would connect, but this created an issue for users with internet service providers that use non-static IP address assignment; NREL resolved this issue by enhancing the firewall for the secure site and removing the IP address restriction).
- Connections between the tools and data are configured for analysis.
- GPS data is spatially interpreted accurately (road map matching routines enhanced during the environment testing process were used to improve spatial-correction of erroneous GPS points).
- The data structure is appropriately documented (through the user manual).
- The data is set up to minimize required user data management to perform custom processing.

The user manual and environment have been structured to permit simple addition of capabilities and descriptions/instructions based on on-going input from new users and analysis efforts.

Figure 5. Navigable user manual in the controlled-access environment

Summary and Future Plans

Increasing numbers of organizations and planning agencies are collecting high-resolution GPS travel data. In spite of the significant effort and expense to collect it, privacy concerns often lead to underutilization of the data. Through the effort described here, NREL and FHWA have partnered to address these concerns, support the needs of numerous data-starved applications, and increase research returns from the original data collection investment. To this end, NREL and FHWA have collaborated with several data providers and users to launch the Transportation Secure Data Center (TSDC).

As described in this paper, the TSDC has been developed and organized with the intent to first and foremost preserve privacy, but to do so in a way that balances security with accessibility and usability of the data for legitimate research needs. The TSDC structure helps support this goal by providing a highly secure enclave for backing up and processing raw data, an easily accessible area for downloading cleansed data, and a controlled-access environment in which approved users can conduct spatial analyses using a variety of tools and reference data.

Future plans for the TSDC include incorporating a process to integrate consistent road grade information into each dataset by using digital elevation models. NREL is currently evaluating how accurately this can be accomplished, along with refining a map-matching and data processing routine to better standardize point-derived calculations (of distance, speed, etc.) across all TSDC datasets. (Figure 6 illustrates an example map-matching result). Other future plans include expanding the number of datasets hosted in the TSDC, incorporating commercial vehicle data and

responding to additional feature requests from the growing user base. One addition may be to create a user-friendly interface for FOSS tools to leverage their speed and cost benefits for users to perform common analyses.

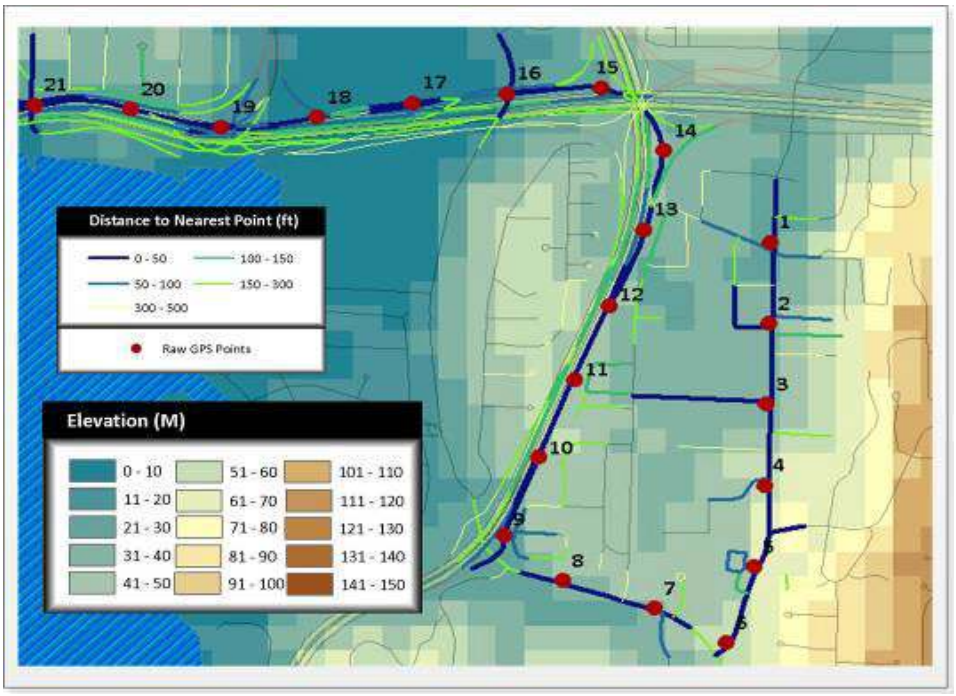

Figure 6. Example map matching output with road reduction filtering and grade adjustment

References

[1] Murakami, E.; Wagner, D. "Can using Global Positioning System (GPS) Improve Trip Reporting?" *Transportation Research C* 7(1999) 149–165.

[2] Gonder, J.; Markel, T.; Thornton, M.; Simpson, A. "Using Global Positioning System Travel Data to Assess Real-World Energy Use of Plug-In Hybrid Electric Vehicles." Transportation Research Record (TRR), *Journal of the Transportation Research Board (TRB)*; No. 2017, Sustainability, Energy and Alternative Fuels 2007; p. 26.

[3] Barth, M.; Boriboonsomsin, K. "Real-World CO_2 Impacts of Traffic Congestion." *Paper #08-2860. Proceedings of the TRB 87th Annual Meeting*; January 2008, Washington, DC.

[4] Tate, E.; Harpster, M.; Savagian, P. "The Electrification of the Automobile: From Conventional Hybrid, to Plug-in Hybrids, to Extended-Range Electric Vehicles." SAE Publication 2008-01-1315. *Proceedings of SAE Congress 2008*; April 2008, Detroit, MI.

[5] Earleywine, M; Gonder, J.; Markel, T.; Thornton, M. "Simulated Fuel Economy and Performance of Advanced Hybrid Electric and Plug-in Hybrid Electric Vehicles Using In-Use Travel Profiles." *Proceedings of the 6th IEEE Vehicle Power and Propulsion Conference (VPPC)*; September 1–3, 2010, Lille, France.

[6] University of Minnesota, Metropolitan Travel Survey Archive website, http://www.surveyarchive.org/. Accessed December 5, 2011.

[7] National Research Council, (2007). *Putting People on the Map: Protecting Confidentiality with Linked Social-Spatial Data.* Panel on Confidentiality Issues Arising from the Integration of Remotely Sensed and Self-Identifying Data. M.P. Gutmann and P.C. Stern, Eds. Committee on the Human Dimensions of Global Change.

Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. http://books.nap.edu/openbook.php?record_id=11865

[8] National Renewable Energy Laboratory, Fuel Cell Vehicle Learning Demonstration website, http://www.nrel.gov/hydrogen/proj_learning_demo.html. Accessed December 5, 2011.

[9] National Renewable Energy Laboratory, Alternative Fuels and Advanced Vehicles Data Center (AFDC) website, http://www.eere.energy.gov/afdc. Accessed December 5, 2011.

[10] U.S. Census Bureau, Center for Economic Studies, Research Data Center Program website, http://www.ces.census.gov/index.php/ces/researchprogram. Accessed December 5, 2011.

[11] NORC data enclave website, http://www.dataenclave.org/index.php/home/welcome. Accessed December 5, 2011.

[12] Khan, M.; Kockelman, K. "Predicting the Market Potential of Plug-In Electric Vehicles Using Multiday GPS Data" Submitted for presentation at the 91st Annual Meeting of the Transportation Research Board in Washington, D.C, January 2012.

[13] Bosley, J.; Eltinge, J.; Fox, J.; Fricker, S. "Conceptual and Practical Issues in the Statistical Design and Analysis of Usability Tests." *Proceedings of the 2003 Federal Committee on Statistical Methodology (FCSM) Research Conference*. http://www.fcsm.gov/03papers/Bosley.pdf