LINKING GPS DATA TO GIS DATABASES IN NATURALISTIC STUDIES: EXAMPLES FROM DRIVERS WITH OBSTRUCTIVE SLEEP APNEA

Jeffrey D. Dawson,[1] Lixi Yu[1], Kelly Sewell[2], Adam Skibbe[2], Nazan S. Aksan,[3] Jon Tippin,[3] Matthew Rizzo,[4]
[1]Dept. of Biostatistics, Univ. of Iowa College of Public Health
Iowa City, Iowa, USA
[2]Dept. of Geographical and Sustainability Sciences,
Univ. of Iowa College of Liberal Arts and Sciences
Iowa City, Iowa, USA
[3]Dept. of Neurology, University of Iowa College of Medicine
Iowa City, Iowa, USA
[4]Dept. of Neurological Sciences, College of Medicine, University of Nebraska Medical Center
Omaha, Nebraska, USA
E-mail: jeffrey-dawson@uiowa.edu

Summary: In naturalistic studies, it is vital to give appropriate context when analyzing driving behaviors. Such contextualization can help address the hypotheses that explore a) how drivers perform within specific types of environment (e.g., road types, speed limits, etc.), and b) how often drivers are exposed to such specific environments. In order to perform this contextualization in an automated fashion, we are using Global Positioning System (GPS) data obtained at 1 Hz and merging this with Geographic Information Systems (GIS) databases maintained by the Iowa Department of Transportation (DOT). In this paper, we demonstrate our methods of doing this based on data from 43 drivers with obstructive sleep apnea (OSA). We also use maps from GIS software to illustrate how information can be displayed at the individual drive or day level, and we provide examples of some of the challenges that still need to be addressed.

INTRODUCTION

Naturalistic driving studies employ electronic sensors and/or video recordings to monitor participants as they operate their vehicles under everyday conditions. In studies of impaired drivers, participants who are aware of their impairments may modify their behavior by limiting driving to essential trips and/or low-risk situations. Therefore, it is important to provide context when analyzing naturalistic driving data. Ideally, this contextualization could be done in an automated fashion, without having to make labor-intensive determinations based on video data. Such contextualization can help separate out the issues of a) how drivers perform within certain driving environments (e.g., road type, speed limit, weather conditions, etc.), and b) how often drivers permit themselves to be exposed to such specific environments.

Drivers with obstructive sleep apnea (OSA), as a group, have higher risk of motor vehicle accidents than drivers without the disorder (Tregear et al, 2009). However, the relationship is complicated by the variability in OSA severity, treatment compliance, and self-awareness of sleepiness (Engleman et al, 1997). We have been conducting a naturalistic study of drivers with OSA, and one of our goals is to compare their driving abilities and exposure strategies to those of

drivers without OSA. We also plan to do correlational analyses within the OSA group, to see how cognitive factors, measures of sleepiness, and treatment compliance relate to driving performance and strategies. Before being able to address these research goals, we need methods to provide context to their electronic driving data. In this report, we outline one such method, which is to link Global Positioning System (GPS) data to Geographic Information System (GIS) maps. We also provide some examples of the challenges associated with this process.

METHOD

Subjects and study overview

The subjects in this study are a subset from a naturalistic driving study of 75 drivers with OSA and 55 controls, ages 30–60 years. All have at least 10 years of driving experience, use a single car as their primary vehicle (90% of driving time), and drive at least 2 hours or 100 miles/week on average. For this report, we are only using data from 43 OSA drivers who do most of their driving in the state of Iowa, so that we could focus our efforts on obtaining GIS data from a single state. The study was approved by the University of Iowa Institutional Review Board for Human Subjects Protection.

Driving monitoring and initial data preparation

Driving behavior was monitored via electronic, video, and GPS outputs from a state-of-the-art instrumentation package installed in each participant's car (McDonald et al, 2012) over a continuous 3.5-month period, with CPAP treatment beginning approximately two weeks into the 3.5-month period for those with OSA. For each trip, a 10-Hz file was created with information pulled from the OBD2 port and from the accelerometers present in the installed device, and a 1-Hz file was created with GPS coordinate information. These two types of files were merged together into one large comma-separated-value (CSV) text dataset per trip, and then concatenated together into one dataset per subject. The amount of data collected, in terms of number of days, number of trips, rows of data, and data file size, varied from subject to subject. One "typical" driver (i.e., who provided a near-median amount of data) took 525 trips on 91 days, resulting in a 587-Mb dataset containing 25 columns (variables) and 3.4 million rows of 10-Hz information.

To facilitate reading the data into GIS software, each subject's 25-variable, 10-Hz CSV file was reduced to a 1-Hz file, with only eight variables. These eight variable included GPS-based coordinates (latitude and longitude), a measure of GPS signal quality, a time/date stamp, and key identification variables. Hence, the number of rows were reduced by 90%, and the number of columns by 68%, resulting in datasets mostly in the 10-40 Mb range. In the case of the above-mentioned "typical" driver, the size of the dataset went from 587 Mb to 21 Mb.

GIS merging and mapping

In preparing to import these data into GIS software and databases, we first eliminated data with null or clearly incorrect GPS information (e.g., coordinates outside of North America). The remaining data were imported into a GIS, specifically ArcGIS 10.1 (ESRI 2013), and converted

to a spatially explicit geodatabase feature. Each data point was given a unique ID based on the driver ID and the observation count fields so it could be consistently referenced and tied back to the source data.

The GIS work was done in two programs: ArcMap (ArcGIS 10.1, ESRI 2013) and ArcGIS Pro (ESRI 2014). For consistency, efficiency, and accuracy, a workflow was created in ArcGIS ModelBuilder and saved as an automation tool. This tool was run on each subsequent dataset.

Road environment data in Iowa are almost exclusively available as centerlines, or line vector data. Attributes that were determined to be of potential impact to a driver's decision-making process or abilities were parsed from the greater datasets. Due to a limit in the precision of our GPS, we "snapped" the GPS points to intersect the closest road centerline. To help with proper extraction of underlying road data values, we buffered each road by 2 feet to provide overlap with the data points. We then extracted the values of each underlying road layer to the GPS point.

The resulting datasets contained 70 new variables pertaining to road culture, including information on speed limit, curves, stop signs, 911-based street names, road surface type, etc. Hence, these datasets were roughly 10 times the size of the eight-variable datasets that were imported into the GIS software. For example, the subject whose data had been reduced to 21 Mb had an outputted dataset that was 186 Mb in size. The Iowa DOT of the state provided data dictionaries to aid in interpreting the values of the field. The processed data were then merged back into the original 25-variable files to be available for future formal analyses.

Statistical analyses

To give a sense of the distribution of the road environment variables, descriptive statistics were calculated for numeric variables across all subjects and rows. Based on preliminary results, we identified several fields whose interpretations remained unclear, which prompted more communications with DOT staff to seek clarification.

RESULTS

Aggregate data

Across the 43 subjects, we amassed a total of 11,121,412 rows (seconds) of data. This corresponds to an average of 258,637 rows per subject, which is approximately 71.8 hours of driving over the 3.5-month period. In Table 1, we show variable names, variable descriptions, and statistical summaries of several numerical variables in the GIS database that we anticipate will be particularly relevant to future studies. Note that there are three variables that should always have positive values (namely, surface width, speed limit, and number of lanes) which have minimum values of 0. This raises the question as to whether these are errors in the GIS database, or whether a 0 should be interpreted as missing values. This issue still needs clarification.

Two variables (not shown) that were deemed by DOT staff to be unreliable were PLANCLASS (a five-level classification for planning purposes, with values such as "Interstate" and "Commercial and Industrial Network") and SUFFSURF (an eight-level ordinal measure of the condition of the road surface). There was also information on other aspects of the driving environment, such as a categorical classification of road type, indicators of rumble strips, street names, terrain type, and zone.

Table 1. Summary of select variables from the GIS database, based on 43 drivers

Variable Name	Variable Description	Mean (SD)	Minimum	Maximum
GRADESTOP	Number of stop signs at grade intersections the segment being traveled	0.73 (1.58)	0.00	15.00
SURFWIDTH	Width of road (in feet)	30.67 (11.92)	0.00	96.00
LIMITMPH	Speed limit (miles per hour)	39.79 (16.24)	0.00	70.00
SLOPE	Transverse slope of a road segment (in percent)	-0.12 (0.54)	-9.00	8.50
NUMLANES	Number of lanes	2.95 (1.39)	0.00	9.00
CURVENUM12	Number of curves ≥28° in segment being traveled	0.34 (0.47)	0.00	2.00

Mapping

We present two kinds of maps to demonstrate how individual data can be viewed and interpreted. In Figure 1, the top 3 panels each show an entire day's driving for one subject, with a rainbow-based color gradient used to show progression throughout the day. In each day, the driver appears to have started (shown in red) somewhere between the intersection of Interstate 80 and Interstate 380, then made one or two loops to the east and/or north before returning. In the lower panel of Figure 1, one can visually compare the patterns across three consecutive days. Such plots could be useful in the future to validate what we are seeing in the GIS data regarding road environment (e.g., higher speed limits when on interstates).

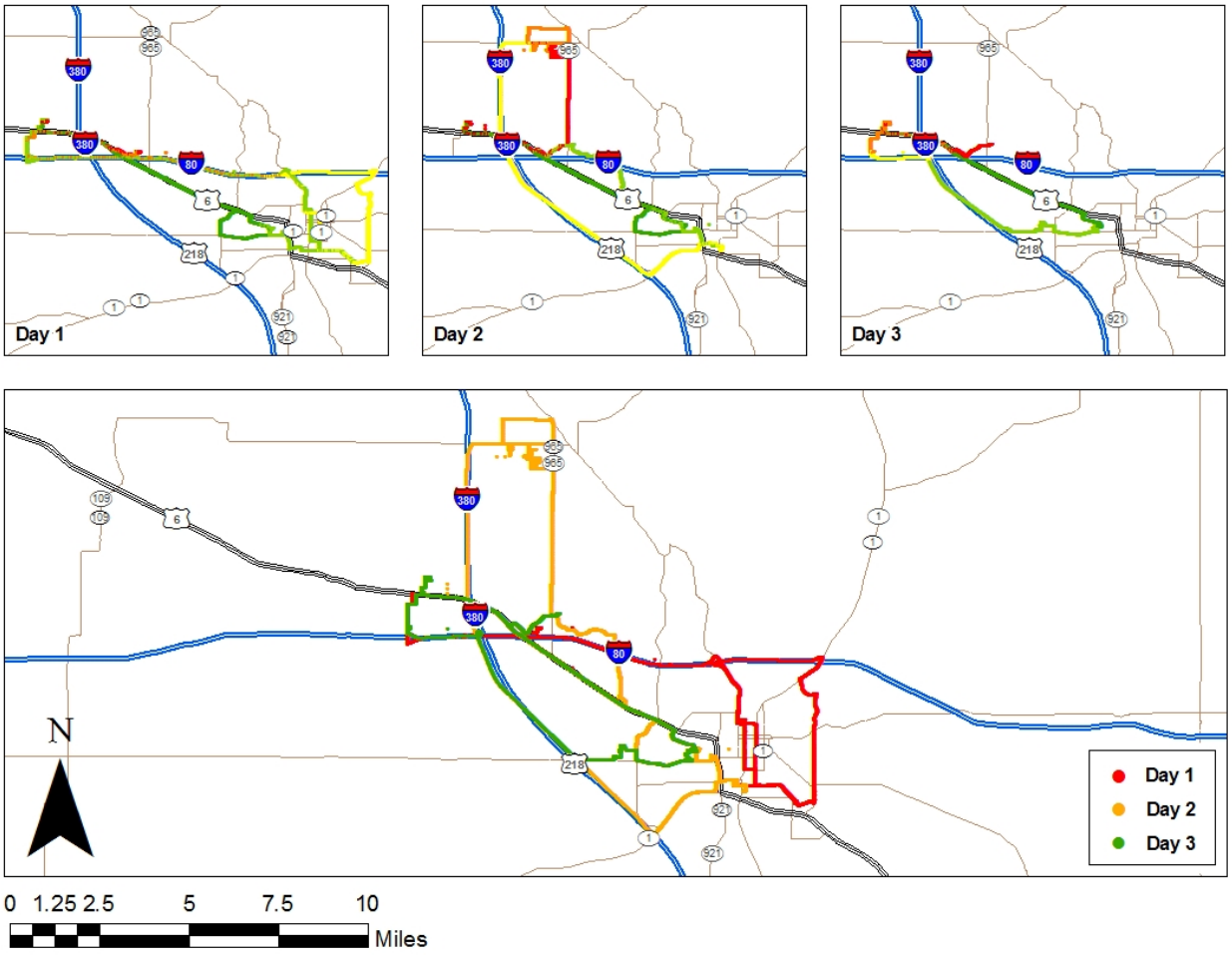

Figure 1. Plots of multiple drives within days (top 3 panels), and drives across days (lower panel)

Figure 2 is similar in nature to the lower panel of Figure 1, but is also magnified and displays satellite images to give additional context. Since this area contains several shopping centers, it appears to be showing driving related to shopping activity for the three days before and after Christmas. Note that this figure also illustrates an important quality control issue. In the lower right-hand quadrant, the subject is driving along the perimeter of a shopping area, and then momentarily disappears in two occasions and incorrectly mapped to other locations: 1) appearing to be on the on-ramp of an interstate (near the center of the map), and 2) appearing to be heading east on the interstate. This shows the lack of precision of the GPS and GIS systems, and shows the need to create algorithms to filter such errors, and to use metrics of driving that will be robust with respect to any errors that are not filtered.

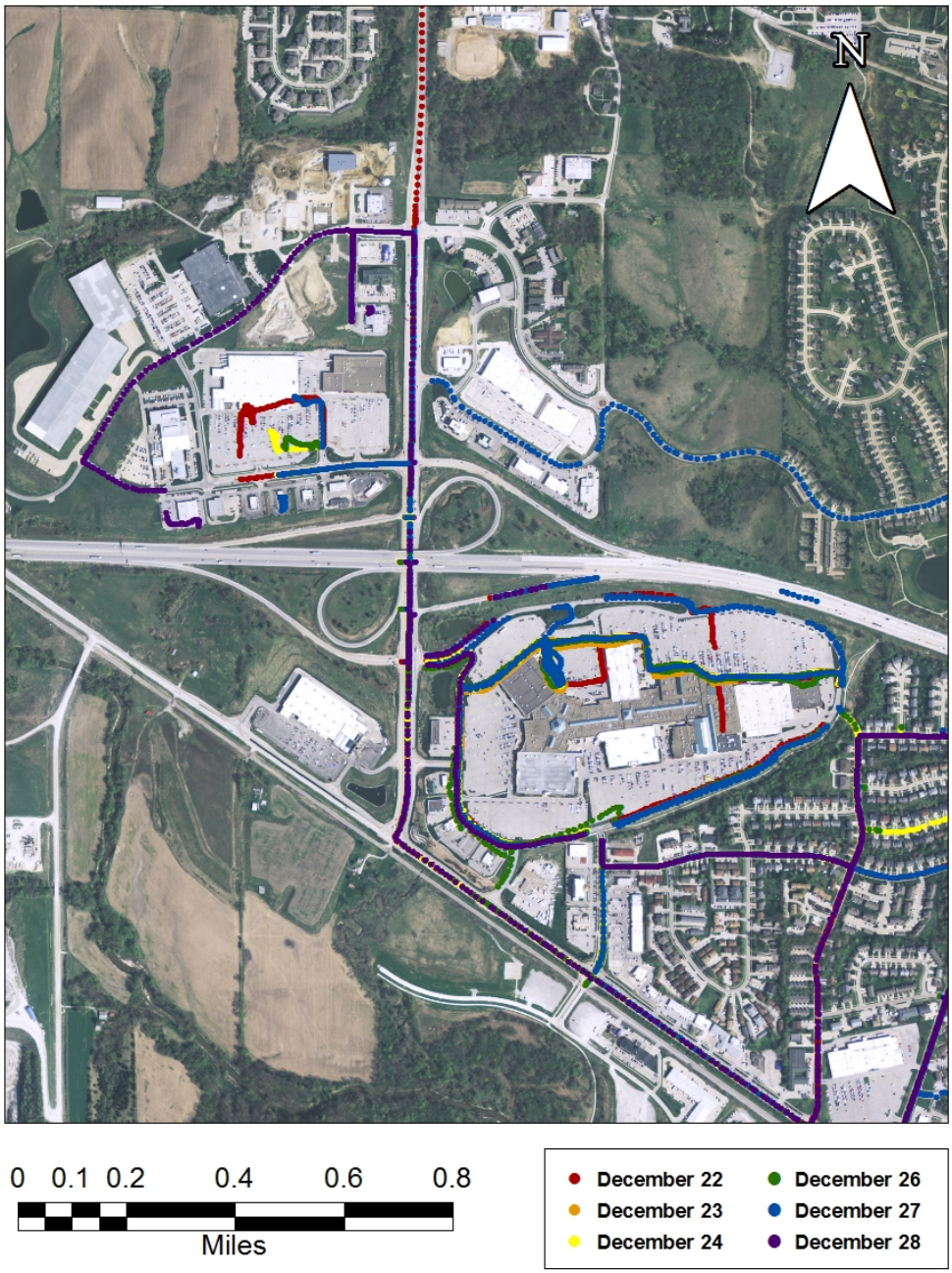

Figure 2. Demonstrating a subject's driving near shopping centers before and after Christmas

DISCUSSION

In this paper, we have described our technique of merging GPS and GIS data together, with the goal of providing better context for analyses based on electronic data. The main idea of the GIS database is that each location in the state is within a specific catchment area, and each catchment area has information on many physical aspects of the road. These catchment areas have varying sizes; hence, larger ones may be picking up information that is not as refined as desired. For example, some catchment areas may include multiple speed limits, stop signs, curves, etc.

Despite such vagueness, the GIS database will be tremendously valuable as we proceed to do formal analyses of our data. For example, we plan to compare lateral accelerometer patterns for drivers with OSA before vs. after treatment, adjusting our analyses for the speed limit of the roads they are using.

Future work includes developing appropriate filters that will improve the accuracy of our contextualization. Although we have demonstrated the usefulness of mapping the data, the hope is that such mapping would only need to be used to validate the environmental data, so that automated analyses can proceed. Validation of GIS data using video clips will also be possible in this study. For example, we have video clips that will enable us to check the accuracy of some of the variables shown in Table 1, such as the number of nearby stop signs, the number of lanes, and the speed limit.

Although we have illustrated our GPS/GIS methods on data from our OSA study, we developed these with the intent of applying them to other studies, as well. For example, we have an ongoing naturalistic study of healthy elderly drivers, which helped support the methodological development described in this paper. It is also our hope that these methods may be considered by other research teams working on analyzing other naturalistic driving studies.

ACKNOWLEDGMENTS

We thank the subjects for their participation and patience. This study was supported by NIH R01 HL091917 and NIH R01 AG017177. We also thank all of our research team for their extraordinary efforts in conducting this study.

REFERENCES

Engleman, H.M., Hirst, W.S.J., & Douglas, N.J. (1997). Under reporting of sleepiness and driving impairment in patients with sleep apnea/hypopnea syndrome. *Journal of Sleep Research*, 6, 272-275.

McDonald, A.D., Lee, J.D., Aksan, N.S., Rizzo, M., Dawson, J.D., & Tippin, J. (2012). Making Naturalistic Driving Data SAX-y. *Proceedings of the VTTI Third International Symposium on Naturalistic Driving Research*, Blacksburg, Virginia.

Tregear S, Reston J, Schoelles K, & Phillips B. (2009). Obstructive sleep apnea and risk of motor vehicle crash: systematic review and meta-analysis. *J Clin Sleep Med*, *5*(6), 573-581.