Executive Summary

In 2010 there were 32,885 traffic fatalities with an estimated 2.24 million people injured in traffic related accidents in the United States (NHTSA). Traffic crashes are one of the leading preventable causes of death in the U.S., with traffic related fatalities accounting for almost 20 percent of all injury deaths (Mokdad, Marks, Stroup, & Gerberding, 2004; CDC).

This application uses crash prediction models for urbanized areas based on county crash data from 48 states and vehicle miles traveled (VMT) and other data for 441 urbanized areas. Crash data are for years 2008 to 2011. VMT and other urbanized area data are for 2010.

Since crash data are for counties and other data are for urbanized areas, we first needed to “crosswalk between urbanized areas and counties” to determine which counties make up which urbanized areas. There is a tradeoff between achieving close match between counties and urbanized areas, and retaining the largest possible sample of county-urbanized area combinations. We have sought a balance.

To develop the crash prediction models, this application uses multiple linear regression analysis (ordinary least square regression, OLS). The dependent variables are total crash rates, injury crash rates, and fatal crash rates. The independent variables consist of VMT per capita as the main predictor of crash rates, and a set of control variables. Using OLS, both linear and logarithmic forms of models are tested to identify the models with the most significant variables and highest predictive power.

The results of the analysis show that VMT per capita is the most significant predictor of total, injury, and fatal crash rates. It is always positively related to crashes. In addition, in the model for total crashes, intersection density has a significant, negative relationship to the crash rate, and employment density has a significant, positive relationship. In the model of fatal crashes, the percentage of 4-way intersections has a significant, positive relationship to the crash rate, and employment density has a significant, negative relationship. Elasticities of crash rates with respect to predictor variables are computed, and in all cases are less than 1.0 in absolute value. That is, crash rates are inelastic with respect to VMT per capita and other variables.

Due to the large unexplained variance in crash rates (low R-squared values), ET+ uses pivot point models to predict crash rates in terms of VMT per capita. Pivot point models are common in sketch planning applications. A pivot point model pivots around the average value of the dependent variable, moving up or down from the average based on the elasticity of the dependent variable with respect to each independent variable. For example, if the elasticity of the total crash rate with respect to VMT per capita is 0.54, a development that has a VMT per capita 10 percent below the county average value would be expected to have a total crash rate 5.4 percent (10%*0.54) below its county average value.
Introduction

Worldwide roughly 1.2 million transportation-related fatalities and fifty million traffic-related injuries occur each year (Peden et al. 2004). These numbers are projected to increase by 65 percent by 2020, thus raising traffic fatalities to the sixth leading preventable cause of death worldwide (World Health Organization 2004). In the United States, we suffer more than 30,000 fatal crashes and an additional 800,000 injuries each year. Motor vehicle traffic deaths remain the leading cause of death among Americans between four and 27 years old (National Highway Traffic Safety Administration 2011). While the United States once had the second-safest transportation system in the developed world, its traffic safety record has fallen behind other developed countries, including England, Australia, and the entirety of the continental Europe. Only if we understand the basic causes of traffic fatalities can we devise policies to reduce their numbers (Methorst & Walker 2010).

Conceptual Framework

A conceptual framework for this app is presented in Figure 1. The published literature is generally consistent with this framework. In this framework, the built environment affects crash frequency and severity through the mediators of traffic volume and traffic speed. Development patterns impact safety primarily through the traffic volumes they generate, and secondarily through the speeds they encourage. Roadway designs impact safety primarily through the traffic speeds they allow, and secondarily through the traffic volumes they generate. Traffic volumes in turn are the primary determinants of crash frequency, while traffic speeds are the primary determinants of crash severity.

Figure 1. Conceptual Framework Linking the Built Environment to Traffic Safety
Traffic Volumes

A key tenet in traffic safety is that humans are prone to error. Failure to notice a potential hazard, delayed response to a perceived hazard, or unexpected behaviors by other road users can all produce traffic crashes. Thus, each and every trip—whether as a motorist, pedestrian, or bicyclist—involves an element of risk.

*Ceteris paribus*, the more vehicular travel, the more risk of crashes. Litman and Fitzroy (2005) examined the relationship between per capita traffic fatalities and vehicle miles traveled (VMT) for urban and rural areas in the United States. As shown in Figure 2, the relationship is roughly linear: as VMT increases, so do traffic fatalities. For urban areas, each 1% increase in travel is associated with a 1% increase in traffic fatalities. For rural areas, each 1% increase in VMT is associated with a 1.5% increase in traffic fatalities (Litman and Fitzroy 2005).

Figure 2: Traffic Fatalities and VMT for Urban and Rural Areas (Litman and Fitzroy 2005)

Balkin and Ord (2001) found that fatalities along individual highway facilities vary seasonally, with crashes increasing during periods that experience seasonal increases in VMT. Conversely, reductions in annual mileage during economic recessions often reduce per capita crash rates. A study of young drivers found that “the consistently significant factor influencing risk of motor vehicle crash involvement was quantity of kilometres driven” (Bath 1993). Similarly, the lower crash rate observed for female drivers is approximately equal to their lower average driving mileage (Butler 1996).
Other studies finding significant relationships between average daily traffic or VMT and crash frequency include Levine et al. (1995a, 1995b), Roberts et al. (1995), Hadayeghi et al. (2003), Lovegrove et al. (2006), and Hess et al. (2004).

**Traffic Speeds**

The other main mediating factor is traffic speed. Simple physics tells us that higher operating speeds give drivers less time to react to unforeseen hazards and result in increased force of impact when crashes occur. At a running speed of 40 mph, a typical driver needs more than 80 feet to stop on wet pavement; at 30 mph, emergency stopping distance drops to just over 40 feet and at 20 mph, it is about 20 feet (see Figure 3).

Figure 3. Typical Emergency Stopping Distance on Wet Pavement for Various Running Speeds (Transportation Research Institute 1997)

Beyond the generalized safety benefits associated with lower vehicle operating speeds, lower speeds have a profound effect on pedestrian safety. Struck by a vehicle traveling 40 mph, a pedestrian has an 85 percent chance of being killed. The fatality rate drops to 45 percent at 30 mph and to 5 percent at 20 mph or less (U.K. Department for Transport 1997; Zegeer et al. 2002a). This relationship is non-linear as well, with crash severity increasing exponentially with vehicle speed (see Figure 4).

Figure 4. Pedestrian Fatality Rates for Collisions at Different Speeds (Zegeer et al. 2002a, p. 13).
Yet perhaps more importantly, the very likelihood that a pedestrian-related crash will occur appears to increase with vehicle operating speeds. In general, low speed, “main street” type designs experience the lowest rates of vehicle-pedestrian crashes, while downtown areas with wide travel lanes and higher operating speeds experience the highest rates (Garder 2004).

**Traffic Conflicts**

It is not traffic speed alone that causes crashes. Rather it is speed differentials among vehicles in the traffic stream. Likewise, it is not traffic volume alone that causes crashes, but rather conflicting movements when traffic volumes are high. The independent role of conflicts comes up in discussions of on-street parking, access management, traffic calming, intersection control, and pedestrian countermeasures. To make this point explicit, an extra box, representing the mediating effect of traffic conflicts, has been added to Figure 1.

**Literature Review**

**Crash Prediction Models (for Facilities)**

For decades researchers have been investigating vehicle crashes and the factors that lead to them (Lord and Mannering 2002). Crash prediction models (CPMs) are used by researchers, transportation, and planning organizations to analyze safety risks and program traffic safety improvements.

Modeling tools have been created by multiple parties at varying levels of detail, but typically at the facility design level. Appendix C of NCHRP report 546, Incorporating Safety into Long-Range Transportation Planning (2006), breaks down available modeling tools and their application. The majority incorporate a reactive analysis to post crash data. From the report, we have selected predictive models available:

- Arizona Local Government Safety Project Analysis Model (LGSP)
- Interactive Highway Safety Design Model (IHSDM)
- Multimodal Transportation Planning Tool (MTPT) GDOT
• Roadside Safety Analysis Program
• SafeNET

The dependent variables in facility crash prediction models use measures such as crash frequency and crash density. The independent variables include average daily traffic, population density, environmental conditions, and roadway design characteristics.

**Crash Prediction Models (for Small Areas)**

Among the tools listed in NCHRP 546, only one predicts crashes at geographic planning unit level. PLANSAFE software uses facility data for a traffic analysis zone to predict several types of crashes. Scenarios may be created by altering transportation and land use outcomes to generate fatal, pedestrian, total, and other crash predictions. The model requires GIS analysis to extract data that are run through a linear regression model with log transformation of the dependent variable.

Dependent variables in small area crash prediction models may be such measures as crash frequency and crash density. Independent variables include: roadway characteristics, transit facilities, bike lanes, population and VMT, crash data, socio-economic and demographic data. Models must then be tested for goodness of fit for predicting outcomes.

Dumbaugh and Li (2011) examined many characteristics of the built environment and correlated them with the number of collisions involving pedestrians, cyclists, and motorists. They found major crash determinants include the total miles of arterial roadways and the presence of strip commercial uses and big box stores. On the other hand, pedestrian-scaled retail uses were associated with lower crash rates. “Each additional mile of arterial thoroughfare was associated with a 9.3% increase in motorist-pedestrian crashes, each additional strip commercial use was associated with a 3% increase in vehicle-pedestrian crashes, and each big box store was associated with an 8.7% increase in vehicle-pedestrian crashes. Four-leg intersections were associated with a 0.9% increase in this crash type.”

Marshall and Garrick (2011) analyzed 230,000 crashes occurring over 11 years in 24 cities in California to determine associations between crashes and street network characteristics, including street network density and street connectivity. Increasing street connectivity - normally associated with street grids - led to an increase in automobile crashes. The authors hypothesized that increased street connectivity leads to increased traffic conflicts and hence more crashes. On the other hand, the severity of crashes, and incidence of fatal crashes, was lower in downtown areas despite their grids. The authors argued a decrease in fatal crashes is a result of lower vehicle speeds on streets in downtown areas.

**Crash Incidence Models (for Larger Areas)**

The literature is replete with studies showing that areas with more residents, more employment, and more arterial lane miles experience more crashes (Levine et al. 1995a,
Such studies may be useful for crash prediction on individual facilities or in small areas. However, they do not explain the relative risk of crashes or the rate of crashes per capita, only overall crash frequency on specific facilities or specific small areas. Where there are more people and jobs, there tends to be more of everything, from traffic to crime to coffee shops.

The crash on a downtown street is as likely to involve someone commuting in from the suburbs as it is a city resident, who may be walking or taking transit to their job or at least driving a shorter distance. The high-volume city street gets blamed for the crash, but the real culprit is the long commute with crash exposure the entire distance.

The better alternative for purposes of development impact assessment is an areawide crash prediction model. Such a model predicts crashes for larger areas, usually including points of origin and destination. If the area is large enough, as with a county, the trip is likely to have been produced and attracted within the area, and the crash rate is likely to reflect the amount of driving residents do.

Given the direct relationship between vehicle miles traveled (VMT) and crash exposure, development patterns that generate lower VMT should also have lower traffic crash rates. If the relationship between VMT and crashes is near-linear, then “sprawling” environments, which are known to generate higher per capita VMT, should also report higher crash rates (Ewing et al. 2002).

In their 2003 article on sprawl vs. traffic fatalities, Ewing et al. (2003b) found that for every 1 percent increase in a county compactness index, all-mode traffic fatality rates fell by 1.49 percent and pedestrian fatality rates fell by 1.47 percent, after adjusting for pedestrian exposure. County compactness was measured in term of population density and street block size. Fatality rates were estimated from the Fatality Analysis Report System (FARS) of the National Highway Traffic Safety Administration. High population densities and small blocks may lead to lower travel speeds, which reduce the severity of crashes.

**Hypothesis**

We would expect compact development patterns to produce fewer traffic accidents and injury accidents, just as they produce fewer fatalities. This is due to reduced vehicle miles driven and possibly also due to lower speeds of vehicle travel. At the same time, compact areas may have more fender benders due stop-and-go driving, even as they have fewer serious crashes due to lower travel speeds. So a priori, our only expectation is that the relationship between sprawl and crash rates will be weaker for total crashes than for fatal crashes. Injury crashes represent an intermediate case, where anything is possible. A priori, the relationship between the built environment and crash rates should be weaker for injury crashes than for total crashes, but stronger than total crashes.
Methods

Our goal is to model crashes in terms of outputs of ET. ET itself produces socioeconomic outputs for different planning scenarios. The ET+ household travel app produces travel outcomes, including VMT estimates. VMT is a logical predictor of crashes, as it relates directly to crash exposure and risk. Built environmental characteristics may also influence crash incidence, beyond the effect of VMT. But how to relate VMT and built environmental characteristics to crashes?

Crash Data

The first thing we needed was crash data. There is no national source of crash data comparable to FARS database of traffic fatalities. Instead, each state, through its Department of Transportation or Public Safety, maintains a comprehensive database of crashes. Crashes are reported if they result in a vehicle being towed away, personal injury, or fatalities. Individual states establish their own reporting thresholds. Crash data are available down to the county level, but typically not for any other geography.

We sought crash data from all 50 states and the District of Columbia. Crash data were obtained from 48 states collected via online databases or per an email/phone request. The survey years ranged from 2008 to 2011 with the majority between 2010 and 2011.

The individual state crash data were compiled into a national database that includes over 5.6 million crashes, 1.7 million injury crashes, and 25,000 fatal crashes. Only 30 percent of crashes resulted in injury, and only 0.4 percent resulted in a fatality.

Table 1. Crash Database

<table>
<thead>
<tr>
<th>fips</th>
<th>State</th>
<th>Abbreviation</th>
<th>All Crashes</th>
<th>Injury Crashes</th>
<th>Fatal Crashes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alabama</td>
<td>AL</td>
<td>124,258</td>
<td>26,943</td>
<td>814</td>
</tr>
<tr>
<td>2</td>
<td>Alaska</td>
<td>AK</td>
<td>12,455</td>
<td>3,657</td>
<td>51</td>
</tr>
<tr>
<td>4</td>
<td>Arizona</td>
<td>AZ</td>
<td>98,033</td>
<td>31,549</td>
<td>612</td>
</tr>
<tr>
<td>5</td>
<td>Arkansas</td>
<td>AR</td>
<td>43,940</td>
<td>12,592</td>
<td>252</td>
</tr>
<tr>
<td>6</td>
<td>California</td>
<td>CA</td>
<td>416,490</td>
<td>161,094</td>
<td>2,520</td>
</tr>
<tr>
<td>8</td>
<td>Colorado</td>
<td>CO</td>
<td>86,558</td>
<td>8,042</td>
<td>283</td>
</tr>
<tr>
<td>9</td>
<td>Connecticut</td>
<td>CT</td>
<td>94,494</td>
<td>23,774</td>
<td>257</td>
</tr>
<tr>
<td>10</td>
<td>Delaware</td>
<td>DE</td>
<td>16,892</td>
<td>4,101</td>
<td>61</td>
</tr>
<tr>
<td>11</td>
<td>District of Columbia</td>
<td>DC</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>Florida</td>
<td>FL</td>
<td>359,796</td>
<td>224,161</td>
<td>2,687</td>
</tr>
<tr>
<td>13</td>
<td>Georgia</td>
<td>GA</td>
<td>306,174</td>
<td>77,150</td>
<td>1,342</td>
</tr>
<tr>
<td>16</td>
<td>Idaho</td>
<td>ID</td>
<td>21,410</td>
<td>8,036</td>
<td>163</td>
</tr>
<tr>
<td></td>
<td>State</td>
<td>Abbreviation</td>
<td>Total Other</td>
<td>Fatalities</td>
<td>Injuries</td>
</tr>
<tr>
<td>----</td>
<td>---------------</td>
<td>--------------</td>
<td>-------------</td>
<td>------------</td>
<td>----------</td>
</tr>
<tr>
<td>17</td>
<td>Illinois</td>
<td>IL</td>
<td>281,878</td>
<td>60,057</td>
<td>835</td>
</tr>
<tr>
<td>18</td>
<td>Indiana</td>
<td>IN</td>
<td>32,337</td>
<td>3,405</td>
<td>37</td>
</tr>
<tr>
<td>19</td>
<td>Iowa</td>
<td>IA</td>
<td>32,055</td>
<td>10,178</td>
<td>150</td>
</tr>
<tr>
<td>20</td>
<td>Kansas</td>
<td>KS</td>
<td>59,740</td>
<td>13,325</td>
<td>354</td>
</tr>
<tr>
<td>21</td>
<td>Kentucky</td>
<td>KY</td>
<td>81,816</td>
<td>14,370</td>
<td>251</td>
</tr>
<tr>
<td>22</td>
<td>Louisiana</td>
<td>LA</td>
<td>122,845</td>
<td>33,971</td>
<td>377</td>
</tr>
<tr>
<td>23</td>
<td>Maine</td>
<td>ME</td>
<td>20,162</td>
<td>4,941</td>
<td>77</td>
</tr>
<tr>
<td>24</td>
<td>Maryland</td>
<td>MD</td>
<td>89,985</td>
<td>30,414</td>
<td>457</td>
</tr>
<tr>
<td>25</td>
<td>Massachusetts</td>
<td>MA</td>
<td>115,378</td>
<td>30,246</td>
<td>331</td>
</tr>
<tr>
<td>26</td>
<td>Michigan</td>
<td>MI</td>
<td>284,049</td>
<td>52,487</td>
<td>834</td>
</tr>
<tr>
<td>27</td>
<td>Minnesota</td>
<td>MN</td>
<td>57,122</td>
<td>16,801</td>
<td>179</td>
</tr>
<tr>
<td>28</td>
<td>Mississippi</td>
<td>MS</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>29</td>
<td>Missouri</td>
<td>MO</td>
<td>141,615</td>
<td>35,279</td>
<td>716</td>
</tr>
<tr>
<td>30</td>
<td>Montana</td>
<td>MT</td>
<td>8,138</td>
<td>1,973</td>
<td>45</td>
</tr>
<tr>
<td>31</td>
<td>Nebraska</td>
<td>NE</td>
<td>29,735</td>
<td>6,519</td>
<td>172</td>
</tr>
<tr>
<td>32</td>
<td>Nevada</td>
<td>NV</td>
<td>50,461</td>
<td>18,744</td>
<td>220</td>
</tr>
<tr>
<td>33</td>
<td>New Hampshire</td>
<td>NH</td>
<td>31,512</td>
<td>6,165</td>
<td>39</td>
</tr>
<tr>
<td>34</td>
<td>New Jersey</td>
<td>NJ</td>
<td>293,595</td>
<td>64,345</td>
<td>573</td>
</tr>
<tr>
<td>35</td>
<td>New Mexico</td>
<td>NM</td>
<td>32,062</td>
<td>9,308</td>
<td>139</td>
</tr>
<tr>
<td>36</td>
<td>New York</td>
<td>NY</td>
<td>439,660</td>
<td>131,131</td>
<td>1,097</td>
</tr>
<tr>
<td>37</td>
<td>North Carolina</td>
<td>NC</td>
<td>208,509</td>
<td>67,983</td>
<td>1,122</td>
</tr>
<tr>
<td>38</td>
<td>North Dakota</td>
<td>ND</td>
<td>7,975</td>
<td>1,752</td>
<td>19</td>
</tr>
<tr>
<td>39</td>
<td>Ohio</td>
<td>OH</td>
<td>296,170</td>
<td>73,427</td>
<td>941</td>
</tr>
<tr>
<td>40</td>
<td>Oklahoma</td>
<td>OK</td>
<td>68,701</td>
<td>23,683</td>
<td>462</td>
</tr>
<tr>
<td>41</td>
<td>Oregon</td>
<td>OR</td>
<td>49,053</td>
<td>23,887</td>
<td>310</td>
</tr>
<tr>
<td>42</td>
<td>Pennsylvania</td>
<td>PA</td>
<td>108,929</td>
<td>48,902</td>
<td>1,191</td>
</tr>
<tr>
<td>43</td>
<td>Rhode Island</td>
<td>RI</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>44</td>
<td>South Carolina</td>
<td>SC</td>
<td>107,673</td>
<td>31,152</td>
<td>700</td>
</tr>
<tr>
<td>45</td>
<td>South Dakota</td>
<td>SD</td>
<td>7,457</td>
<td>2,195</td>
<td>26</td>
</tr>
<tr>
<td>46</td>
<td>Tennessee</td>
<td>TN</td>
<td>195,799</td>
<td>48,293</td>
<td>903</td>
</tr>
<tr>
<td>47</td>
<td>Texas</td>
<td>TX</td>
<td>430,226</td>
<td>143,142</td>
<td>2,818</td>
</tr>
<tr>
<td>48</td>
<td>Utah</td>
<td>UT</td>
<td>46,272</td>
<td>14,153</td>
<td>217</td>
</tr>
<tr>
<td>49</td>
<td>Vermont</td>
<td>VT</td>
<td>5,116</td>
<td>734</td>
<td>13</td>
</tr>
<tr>
<td>50</td>
<td>Virginia</td>
<td>VA</td>
<td>102,742</td>
<td>37,544</td>
<td>NA</td>
</tr>
<tr>
<td>51</td>
<td>Washington</td>
<td>WA</td>
<td>98,878</td>
<td>32,725</td>
<td>422</td>
</tr>
<tr>
<td>52</td>
<td>West Virginia</td>
<td>WV</td>
<td>29,946</td>
<td>9,050</td>
<td>166</td>
</tr>
<tr>
<td>53</td>
<td>Wisconsin</td>
<td>WI</td>
<td>112,516</td>
<td>28,965</td>
<td>515</td>
</tr>
<tr>
<td>54</td>
<td>Wyoming</td>
<td>WY</td>
<td>14,112</td>
<td>3,643</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td></td>
<td><strong>5,674,720</strong></td>
<td><strong>1,715,988</strong></td>
<td><strong>25,885</strong></td>
</tr>
</tbody>
</table>
**VMT and Other Data**

While crash data are available for counties, VMT data are available only for urbanized areas. This mismatch creates a challenge. Our dependent variables are county-specific crash rates while our most important independent variable is urbanized area VMT from federal *Highway Statistics*. We needed to relate one to the other. This involved the matching procedure described in the next section.

Another challenge resulted from the fact that the VMT reported in *Highway Statistics* is not for urbanized areas as defined by the U.S. Census but rather for urbanized areas as drawn by metropolitan planning organization and state departments of transportation and approved by the Federal Highway Administration (FHWA). According to FHWA, “the boundaries of the [FHWA urbanized] area shall encompass the entire urbanized area as designated by the U.S. Bureau of the Census plus that adjacent geographical area as agreed upon by local officials in cooperation with the State.” Most researchers use the census boundaries for their analysis and delete urbanized areas from the sample if the census and FHWA boundaries were hugely different. We chose not to make such approximations or lose many cases, and therefore set out to find FHWA adjusted boundaries for urbanized areas in a geospatial shapefile format, which we could then use to conduct spatial analyses in GIS (see Figure 5).

Figure 5. 2000 Census and FHWA-Adjusted Urbanized Areas Boundaries for Atlanta
FHWA advised us to contact individual state DOT offices for their shapefiles, which we did. This sometimes required several calls to find the right office. In this way, we were able to obtain shapefiles for all 50 states and 443 urbanized areas. We then combined the individual state files into one national shapefile by using the “merge” function in GIS. Many of the urbanized areas cross state boundaries and in this case we had more than one polygon for each urbanized area. So, we used the “dissolve” function in GIS to integrate those polygons into one for each urbanized area.

After cleaning the data, we did several spatial joins in GIS to capture data from other sources. For example, we used the “centroid” function to join 2010 census tracts to
FHWA adjusted urbanized areas. We then aggregated values of per capita income for census tracts to obtain urbanized area weighted averages (weighted by population). The initial sample consisted of 443 urbanized areas. When further limited to urbanized areas with crash and other data, our final database consisted of 347 urbanized areas, including nearly all the large urbanized areas in the U.S. and most of the small ones.

Crosswalking between Urbanized Areas and Counties

Since our dependent variables are for one geography—counties—and our independent variables are for another—urbanized areas—it is necessary to crosswalk between them. In the simplest case, an urbanized area falls strictly within one county and forms the urban core of that county. This provides a one-to-one match. Such urbanized areas tend to be small, for example, Colorado Springs and Davis, CA. If the population of the county is only slightly larger than the population of the urbanized area, it may be reasonable to associate county crash rates with urbanized area VMT and built environments. The Colorado Springs urbanized area represents 90 percent of surrounding El Paso County, CO, so here we have a match (Figure 6). In contrast, the Davis urbanized area represents only 36 percent of surrounding Yolo County, CA (Figure 7). Here there is no match.
Figure 6. Colorado Springs Urbanized Area and El Paso County
A second type of relationship exists when an individual urbanized area covers several counties. This provides a one-to-many match. This kind of match describes many of the largest urbanized areas such as Atlanta and Chicago. In these cases, the urbanized area forms the urban core of multiple counties and may represent the greater part of their population. It is thus reasonable to associate the counties’ combined crash rates with the urbanized area VMT and built environment.

A third type of relationship has multiple urbanized areas within a single county. This provides a many-to-one match. One of many urbanized areas is usually much larger than the others. For example, Los Angeles County contains one large urbanized area (Los Angeles–Long Beach–Santa Ana) and three small urbanized areas (Lancaster–Palmdale, Santa Clarita, and Thousand Oaks). In such case the largest urbanized area may or may not represent the bulk of the county, and it may or may not be reasonable to associate county crash rates with the urbanized area’s VMT and built environment.
Finally, an urbanized area may be divided somewhat evenly between two or more counties and represent only a small portion of each county. The Danbury, CT urbanized area is divided among three counties, and contains only 14 percent of their combined population. It would be specious to assign the corresponding county crash rates to this urbanized area.

There is a tradeoff between achieving a close match between counties and urbanized areas, and retaining the largest possible sample of county-urbanized area combinations. Operationally, this is how we made our matches. We first intersected urbanized areas with counties to generate new polygons that uniquely relate urbanized areas, or portions thereof, to counties. For example, the Albuquerque urbanized area is divided between two counties, Bernalillo and Sandoval counties, and two polygons were created from the intersection (see Figure 8). The polygons were then joined with census block groups to obtain population estimates for each polygon. These population estimates were compared to county populations to see if the urbanized area covers a substantial portion of the counties. In the case of Albuquerque, the urbanized area houses 97 percent of Bernalillo County’s population and 81 percent of the Sandoval County’s population.

Figure 8. Albuquerque Urbanized Area and Bernalillo and Sandoval Counties
We used two matching procedures to ensure that urbanized areas were reasonably comparable to component counties. We tested various threshold values to define a match, and finally settled on a 75 percent population standard.

If the population of the urbanized area, or portion thereof, represented more than 75 percent of the county population, the two were treated as a match and county crash statistics and populations were assigned to the urbanized area. Portions of urbanized areas were then re-aggregated with associated attributes of counties, and average crash rates (combined county crashes divided combined county population) were computed. In the case of Albuquerque, the portions in Bernalillo and Sandoval counties were recombined. In this manner, 199 of 436 urbanized areas were wholly or partially matched to their component counties.

The final matching procedure ensured that the portions of urbanized areas thus combined were, indeed, representative of the entire urbanized area. To declare a match, the matched portions had to house more than 75 percent of the urbanized area’s total population. Otherwise, the VMT and other attributes of the urbanized area as a whole might not be representative of the matched portions and their component counties.

One hundred percent of the Albuquerque urbanized area population falls within Bernalillo and Sandoval counties. For Albuquerque, urbanized area VMT, sociodemographic, and built environmental characteristics can be used to model average county crash rates. By contrast, only 67 percent of the population of the Nashville urbanized area is contained within the one county (Davidson County) out of six that achieved a match with the urbanized area (see Figure 9). The urbanized area represents less than 75 percent of the population of the each of the five other counties. For example, the urbanized area encompasses only 33 percent of the population of Wilson County to the east of Davidson County. Nashville was thus lost to our sample. Of the 199 urbanized that matched one or more counties, 172 contained enough population within the matched portions of counties to meet the 75 percent threshold.
Figure 9. Nashville Urbanized Area and Its Component Counties

Variables

Table 2 provides a list of variables used to model crash rates. Our dependent variables relate to counties individually or in groups:

- total crash rate (crashrate)
- non-fatal injury crash rate (injuryrate)
- fatal injury crash rate (fatalrate)

Our independent variables are VMT per capita, sociodemographic variables, and built environmental variables for matched urbanized areas. VMT per capita is our main predictor of crashes. It is a measure of exposure and risk. The sociodemographic
variables, household size and household income, serve as control variables. They are outputs of ET. The built environmental variables cover three dimensions relevant to crash prediction: development density, land use mix, and street accessibility. These dimensions distinguish compact development patterns from sprawling ones, or put another way, highly accessible development patterns from those that are not (Ewing et al. 2002).

Table 2. Variables Used to Model Crash Rates

<table>
<thead>
<tr>
<th>Variables</th>
<th>Data Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variables</strong></td>
<td></td>
</tr>
<tr>
<td>crashrate</td>
<td>traffic crash rate per 100,000 residents</td>
</tr>
<tr>
<td>injuryrate</td>
<td>injury crash rate per 100,000 residents</td>
</tr>
<tr>
<td>fatalrate</td>
<td>fatal crash rate per 100,000 residents</td>
</tr>
<tr>
<td><strong>Independent variables</strong></td>
<td></td>
</tr>
<tr>
<td>vmtcap</td>
<td>vehicle miles per person per day</td>
</tr>
<tr>
<td>hsize</td>
<td>average household size</td>
</tr>
<tr>
<td>hhinc</td>
<td>median household income</td>
</tr>
<tr>
<td>popden</td>
<td>gross population density</td>
</tr>
<tr>
<td>empden</td>
<td>gross employment density</td>
</tr>
<tr>
<td>job-pop</td>
<td>job-population balance</td>
</tr>
<tr>
<td>entropy</td>
<td>employment mix</td>
</tr>
<tr>
<td>intden</td>
<td>intersection density</td>
</tr>
<tr>
<td>int4way</td>
<td>percentage of 4-way intersections</td>
</tr>
</tbody>
</table>

**Analytical Method**

We used multiple linear regression analysis (ordinary least squares regression or OLS) to model crash rates. Using OLS, we tested both linear and logarithmic forms of variables. The linear forms outperformed the logarithmic in terms of the significant levels of variables (t-statistics and probability levels) and the predictive power of models (R2s). Logarithmic transformations are sometimes used to account for nonlinear relationships among variables, reduce the problem of outlying data points, and produce more normally distributed dependent variables. In this case, the linear dependent variables were actually
more normally distributed than the logarithmic variables (see Figures 10 and 11). And nonlinearity and outliers were not problematic.

We also estimated models both with and without constant terms (that is, with and without regression intercepts). With the latter, the regression line or plane is forced through the origin. The rationale for forcing the regression line or plane through the origin is that if all independent variables were 0, including VMT per capita, we would expect there to be no crashes. To choose between these two model forms, we consulted with Professor William Greene of New York University, one of the world’s foremost experts on econometric modeling and a statistical consultant on this project. Bill recommended that we include the constant terms in our model, with the following explanation:

The regression is fit to the data that are within the range of your (the world's) experience. Zero miles traveled is nowhere near that range. It is true that the constant term in your regression is an estimator of zero. But, there are any number of other observed and unobserved influences that will affect the regression results. You can afford to let the data speak and allow the constant term in your regression be nonzero - the cost is a single degree of freedom. I would not impose the restriction. If the constant terms out not to be significant, you can argue that it makes sense that it wouldn't and leave it at that. But, imposing the constraint might significantly degrade the fit. And, it would complicate fit measures such as R-squared that assume that the regression has a constant.

Figure 10. Histogram of Crash Rates for Selected Urbanized Areas (as estimated from component counties – only urbanized areas in final sample)
Results

Our crash incidence models are presented in Tables 3 through 5. The tables contain regression coefficients, standard errors, t-statistics, probability values, and R-squared statistics. The regression coefficients tell us how much crash rates increase with unit increases in independent variables. t-statistics and probability values tell us how statistically significant independent variables are as predictors of crash rates. R-squared statistics, which measure goodness-of-fit, tell us what proportion of the variation in crash rates is explained by the model as a whole.

From Table 3, VMT per capita is directly related to the total crash rate per 100,000 residents and is significant at the 0.001 probability level or beyond. This is as expected. More driving and exposure lead to more crashes. Also highly significant is intersection density, which has an inverse relationship to crashes and is significant at the 0.001 level or beyond. This is not entirely expected, since crashes are known to be concentrated at intersections. However, this result is consistent with studies by Dumbaugh and colleagues (see above) and the literature generally (Ewing and Dumbaugh 2009). It also has a logical explanation. Short blocks and frequent intersections slow traffic, thereby reducing crash risk. Finally, employment density has a less significant, but positive,
relationship to total crashes. Concentrations of employment, particularly in commercial strips, may lead to more traffic conflicts and hence higher crash rates.

The R-squared and adjusted R-squared are low. The model explains about 17 percent of the variation in crash rates. The low R-squared is a problem, handled with the use of a pivot point model. More on this later.

Table 3. Best-Fit Regression Model for Total Crashes per 100,000 Residents (linear form)

<table>
<thead>
<tr>
<th></th>
<th>coeff</th>
<th>std error</th>
<th>t-value</th>
<th>p-value</th>
<th>tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>1811.8</td>
<td>408.0</td>
<td>4.44</td>
<td>&lt; 0.001</td>
<td></td>
</tr>
<tr>
<td>vmtcap</td>
<td>50.9</td>
<td>14.1</td>
<td>3.63</td>
<td>&lt; 0.001</td>
<td>0.94</td>
</tr>
<tr>
<td>empden</td>
<td>0.493</td>
<td>0.230</td>
<td>2.15</td>
<td>0.033</td>
<td>0.42</td>
</tr>
<tr>
<td>intden</td>
<td>-24.7</td>
<td>6.14</td>
<td>-4.03</td>
<td>&lt; 0.001</td>
<td>0.43</td>
</tr>
<tr>
<td>R2</td>
<td>0.169</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>adj R2</td>
<td>0.154</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Table 4, the injury crash model has only one significant variable at the 0.05 probability level (though entropy is significance at the 0.10 level). VMT per capita is significant at the 0.001 level or beyond. More VMT per capita translates into more injury crashes. The R-squared of the model is 0.138, meaning that the model explains about 14 percent of the variation in injury crash rates. This too is a low R2, which will be discussed later.

Table 4. Best-Fit Regression Model for Injury Crashes per 100,000 Residents (linear form)

<table>
<thead>
<tr>
<th></th>
<th>coeff</th>
<th>std error</th>
<th>t-value</th>
<th>p-value</th>
<th>tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>216.4</td>
<td>93.2</td>
<td>2.32</td>
<td>0.021</td>
<td></td>
</tr>
<tr>
<td>vmtcap</td>
<td>17.8</td>
<td>3.93</td>
<td>4.54</td>
<td>&lt; 0.001</td>
<td>NA</td>
</tr>
<tr>
<td>R2</td>
<td>0.110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>adj R2</td>
<td>0.105</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Table 5, the fatal crash rate increases with VMT per capita, declines with increasing employment density, and increases with the percentage of 4-way intersections. The relationship of VMT per capita to fatal crashes is positive, as expected. More VMT per capita translates into more fatal crashes. The relationship of employment density to fatal crashes is negative, which is the reverse of its relationship to total crashes. Upon reflection, this seems reasonable. There may be more fender benders in areas of high employment density, but fewer high-speed, fatal crashes. Finally, the relationship between percentage of 4-way intersections and fatal crashes is positive. This is consistent with the crash literature (Ewing and Dumbaugh 2009). Four-way intersections have more conflict points than 3-way intersections, and also can be traversed at higher speeds.
by cross street traffic. The R-squared of the model, while still low, is a more respectable 0.201.

Table 5. Best-Fit Regression Model for Fatal Crashes per 100,000 Residents (linear form)

<table>
<thead>
<tr>
<th></th>
<th>coeff</th>
<th>std error</th>
<th>t-value</th>
<th>p-value</th>
<th>tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>0.082</td>
<td>2.81</td>
<td>0.029</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>vmtcap</td>
<td>0.319</td>
<td>0.077</td>
<td>4.16</td>
<td>&lt; 0.001</td>
<td>0.78</td>
</tr>
<tr>
<td>empden</td>
<td>-0.003</td>
<td>0.001</td>
<td>-3.94</td>
<td>&lt; 0.001</td>
<td>0.94</td>
</tr>
<tr>
<td>int4way</td>
<td>0.120</td>
<td>0.049</td>
<td>2.45</td>
<td>0.015</td>
<td>0.81</td>
</tr>
<tr>
<td>R2</td>
<td>0.201</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>adj R2</td>
<td>0.186</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Elasticities**

An elasticity is a percentage change in a dependent variable accompanying a 1 percent change in an independent variable. It is a dimensionless measure of effect size commonly used in economics and planning.

From our regression models, we computed elasticities of crash rates with respect to VMT per capita and other significant variables using the elasticity formula:

\[
\text{elasticity} = b \times \frac{\text{x mean}}{\text{y mean}}
\]

where \( b \) is the regression coefficient, \( x \) mean is the average value of the independent variable, and \( y \) mean is the average value of dependent variable. For our different models, elasticity values are shown in Table 6.

Of note is the fact that the crash rate is inelastic with respect to VMT per capita. That is, the crash rate does not rise as fast as VMT per capita rises, or fall as fast and VMT per capita falls. Perhaps, longer trips farther from home (including long commutes) are safer on a vehicle mile basis than short trips close to home. The old adage, most accidents happen close to home, may be literally true. These elasticity values will be used in the pivot point model described in the next section.
Table 6. Elasticities of Crash Rates with Respect of VMT per Capita and Built Environmental Variables

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Elasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Crash Rate</strong></td>
<td></td>
</tr>
<tr>
<td>VMT per Capita</td>
<td>0.54</td>
</tr>
<tr>
<td>Employment Density</td>
<td>0.18</td>
</tr>
<tr>
<td>Intersection Density</td>
<td>-0.53</td>
</tr>
<tr>
<td><strong>Injury Crash Rate</strong></td>
<td></td>
</tr>
<tr>
<td>VMT per Capita</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>Fatal Crash Rate</strong></td>
<td></td>
</tr>
<tr>
<td>VMT per Capita</td>
<td>0.89</td>
</tr>
<tr>
<td>Employment Density</td>
<td>-0.29</td>
</tr>
<tr>
<td>Percentage 4-Way Intersections</td>
<td>0.39</td>
</tr>
</tbody>
</table>

**Pivot Point Models**

In light of the large unexplained variance in our county crash incidence models, we decided to use pivot point models to predict crash rates in terms of VMT per capita, and other significant variables. A pivot point model pivots around the average value of a dependent variable using elasticities of the dependent variable with respect to independent variables and considering differences in values of the independent variables from the average.

For example, the elasticity of the total crash rate with respect to VMT per capita is 0.54. If the total crash rate in a particular county is 2,000 crashes per year per 100,000 residents, the average VMT is 20 vehicle miles per person per day, a particular development will generate 10 vehicle miles per person per day according to ET+, and the development has or will have 1,000 residents, the number of total crashes generated by the development would be predicted to be:

\[ \text{total crashes} = 1,000 \times \left( \frac{2,000}{100,000} \right) \times \left( 1 + 0.54 \times \left( \frac{10-20}{20} \right) \right) = 1.46 \text{ per year} \]

ET+ contains a look-up table with model elasticities and average crash rates by severity for almost every county in the U.S. It is around these rates, and with these elasticities, that ET+ pivots. The look-up table also contains actual county average values of the independent variables in our models (and estimates of county VMT per capita), values which serve as a baseline for the pivots.
Discussion

Several factors give us confidence in these models for crash prediction purposes. First is the consistency of results regardless of where we set population matching thresholds for urbanized areas and counties. Second is the high significance levels of VMT per capita, intersection density, and other significant variables, and their plausible relationships to crash frequency. These models are certainly superior to the standard travel demand model assumption of constant crash rates per 100,000 VMT. Perhaps most important is our use of average county crash rates as pivot points for crash prediction. Rather than use the models themselves to predict crash rates, we estimated elasticities of crash rates from regression coefficients, and now pivot around county average rates using these elasticities. This eliminates much of the potential error associated with residuals (unexplained variance) in our regression equations.

These models also have limitations. The study design is ecological in nature. It treats each county as a unit of homogeneous density, mix, and street accessibility, and assigns to it a single crash rate, even though there are likely to be large differences within its borders.

The explanatory power of our regression models, represented by their R2s, is lower than we might wish. The unexplained variance is not too surprising considering:

- the inherent mismatch between counties and urbanized areas
- the different crash reporting requirements set by different states
- the different years of crash data reported by different states
- the many independent variables missing from our list, including any measure of traffic speed

The crash data studied are based on location of crash, while the population density and street accessibility data are based on place of residence, which may be different. To the extent that crashes occurred during the morning or evening commute, a (reassuring) bias towards the null may exist. In other words, because most commuters who cross county borders live in lower-density bedroom communities and work in higher-density central cities, the crash rate in urban counties would be inflated relative to the population living there. Using these databases, we could not determine the extent to which such bias, if any, existed. One solution would be to study the relationship at the (multi-county) metropolitan area level, but this would be at the expense of desired precision in the measurement of differences within metropolitan areas.

Additional studies are needed to confirm these findings and extend our knowledge in key areas. An exploration of the relationship between vehicle speed, crash rates, and specific street design features common to urban sprawl (e.g., wide, long streets) would help guide countermeasures.
References


Traffic Safety App


Data cited from the Fatality Analysis Reporting System and the General Estimates System.


