

An integrated approach for studying the safety of road networks: logistic regression models between traffic accident occurrence and behavioural, environmental and infrastructure parameters

F. Crocco, S. De Marco & D. W. E. Mongelli

Territorial Planning Department, University of Calabria, Italy

Abstract

The current tendency is to attribute the cause of each road accident to certain factors that, closely related, define a system whose crisis points can construct serious elements of danger for road safety. The situation then becomes complex when it is impossible to attribute the cause of an accident to a sole component of the system 'driver – vehicle – environment – infrastructure', because more than one problem is present when investigating the path leading up to an event and it is unclear in what measure to hold the driver responsible for actions that have risked his or her own life, and in many cases, the lives of others. This is a problem that cannot be solved by assigning closed formulas when the investigation of different locations presents diverse characteristics and the variables in play can change even slightly from case to case in the same network under study. In this article, logistic regression models are proposed to evaluate the reliability of a roadway, where it is understood as a probability that a driver travels along a path in a specific period of time and in specific operative conditions, without an accident with injury or death occurring. The analysis has been conducted globally, placing in contemporaneous correlation infrastructural, environmental and behavioural components. The models have been applied to a stretch of freeway and state roads in the province of Cosenza (Italy) by defining the seriousness of the accidents that occurred.

Keywords: accident, driver, environment.



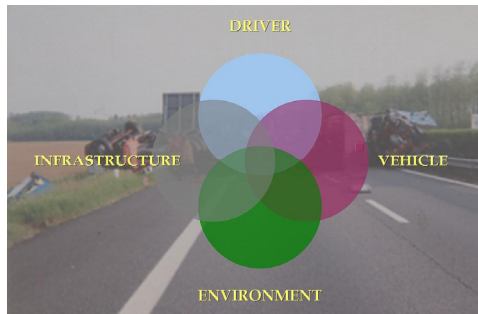


Figure 1: The system “driver – vehicle – environment – infrastructure”.

1 Introduction

Road accident occurrence is an ever-present problem of widespread interest given the repercussions this phenomenon has on the collective, be it in terms of human life loss or social impact. In fact, road safety begins with the interaction of various factors related to the human behavioural environment, infrastructural characteristics, vehicle conditions and the surrounding environment. As a result, it is extremely difficult to unequivocally pinpoint incidents to a specific defective field or aspect.

Every movement can be seen as part of a system where the components consist of drivers, vehicles, the environment and the infrastructure which, if analyzed separately, have to be investigated in the complexity of their own interactions. The moment in which one or all of the components manifest functional anomalies, dangerous situations of varying levels arise that can determine the total breakdown of a system.

Each component can be distinguished by diverse risk factors and keeping this in mind, the strategies to attain the objectives for a safe road can also be defined.

Road examination requires careful attention to the geometry axis or section; the surface characteristics of pavement, safety tools (guard rails, road signs, lighting, etc.); test criteria of the passive safety of roads; regulatory systems and traffic control; traffic accident statistics, road rules of conduct, etc.; traffic and infrastructure service levels. Regarding environment, traffic conditions cited include: fog, ice, constant rain etc. Fatal factors for the driver include excessive velocity, non-compliance with road regulations, distracted driving or poor physical driving conditions such as fatigue, alcohol and drug abuse (both legal and illegal), etc. Characteristics among drivers that increase road incidents include: age, sex, feelings of superiority, irritability and social influence.

The unpredictable character of events and the complexity of factors that appear cause the regulation of safety, understood as the identification of reasons for danger and their removal, to be a difficult and complex operation. It is thus the objective of this article to propose an integrated approach for studying road safety that envisions a unitary system in which infrastructural, environmental and behavioural components are taken into consideration at the same time. From this

point of view, the estimated reliability of a road system is translated into the value that each individual mobility component can take on for a specific period of time, in specific infrastructural and environmental conditions, within a complex enough system where, at times, the behavioural component plays an important rule.

The system “driver – vehicle – environment – infrastructure” is always functional until the moment in which breakdowns or accidents occur.

The reliability of the system is non-existent if the interactions of the components it is composed of cause accidents with injury and/or death. In other words, the reliability of the system is evaluated as the probability that the systems functions correctly in a specific period of time and in specific operative conditions (maintenance of infrastructure, traffic, environment and above all the behaviour of the road user).

2 State of the art

The issue of road system reliability has been discussed by many authors that have investigated it in terms of competitive system loss. This can occur on specific occasions in the moment in which some of the variables in use take on heightened values and are incompatible with infrastructure, or even when they have broken some branch of the network.

In recent years, Bayesian methods have found several applications in traffic crash analysis. Christiansen et al. [1] and MacNab [2] developed hierarchical Poisson models for crash counts and surveillance data. Miaou and Song [3] used Bayesian generalized linear mixed models with multivariate spatial random effects to rank sites by crash cost rate. Three crash severities were analyzed: fatal, incapacitating injury and non-incapacitating injury crashes at the county level. The authors were mainly interested in ranking criteria and did not discuss correlation structures.

Liu et al. [4] used a hierarchical Bayesian framework to estimate ZIP regression models and develop safety performance functions (SPFs) for two-lane highways. Pawlovich et al. [5] employed a Bayesian approach to assess impacts of road design measures on crash frequencies and rates. Washington and Oh (2006) developed a Bayesian methodology for incorporating expert judgment in ranking countermeasure effectiveness under uncertainty. Bijleveld [6] proposed a method to estimate the covariances between crash and injury counts from within the accident data, using individual accident information on the relevant outcomes, for example the number of victims and fatalities. This method however, estimates only covariances within an observation ignoring covariances between observations. Song et al. [7] further explored the issue of multivariate spatial models. Instead of correlated severities, the authors studied correlated crash types: intersection crash, intersection related crash, driveway access crash, and non-intersection crash. The authors found that the model with correlated spatial random effects fit the data better than the model with independent random effects. Although summary statistics of the correlation coefficients between the responses were not provided, plots of the posterior distribution of the coefficients



indicated significant correlations ranging from 0.3 to 0.6. Ma and Kockelman [8] used multivariate Poisson regression within a Bayesian framework to estimate the number of victims per road segment in five different severity classes: fatal, disabling injury, non-disabling injury, possible injury and non-injury. The model did not allow for extra-Poisson variation in the data and restricted the covariances for the five different severity levels to be identical and positive. Park and Lord [9] extended the Bayesian Multivariate Poisson specification to include extra-Poisson variation and relaxed the covariance specification to allow for different covariances among severity levels by including a log-Normal random effect. The authors analyzed crash counts for three-legged unsignalized intersections and five severity levels: fatal, incapacitating injury, non-incapacitating injury, minor injury and property damage only. The posterior means of the correlation matrix indicated high correlation between different crash severities (from 0.69 to 0.85). Ma and Kockelman [10] also used a Bayesian Multivariate Poisson log-Normal specification to model crash counts at the segment level. However, the correlations found by the authors were significantly lower than the correlations found in [18], ranging from 0.01 to 0.43.

Aguero-Valverde and Jovanis [11] uses Full Bayes Multivariate Poisson log-Normal models to estimate the expected crash frequency for different crash severity levels and then compares those estimates to the independent or Univariate Poisson log-Normal estimates. The finding of high correlation between contiguous severity levels is consistent with some of the literature, but additional tests of multivariate models are recommended. The improved precision has important implication for identification of sites with promise (SWiPs) as one formulation includes the standard deviation of crash frequencies for similar sites as part of the assessment of SWiPs.

3 The construction of the database

The relevant files are shared among provincial prefectures. At each prefecture, meetings were organized with relevant authority figures (Traffic Police, Carabinieri, Local Police) where the uniformity of the report files were established in order to have an accepted and homogenous observation base.

The information reported in the files is qualitative and concerns:

- weather conditions: calm, cloudy, rain, fog/mist, snow;
- road surface: dry, slippery, snowy, wet, icy;
- state of pavement maintenance: indicated by normal and low-quality;
- geometric characteristics: straight, curve, tunnel, bridge, intersection, exit, incline/decline;
- type of accident: frontal, side, pile-up, road crossover, pedestrian run-over, hitting fixed objects, running off the road.

The data was organized in a database where specific fields indicate for each accident date of accident, time, place, road type, cause of accident, collision type, infrastructure details in which the event occurred (geometric characteristics, pavement status), weather condition information, driver information (age, sex,

seatbelt/helmet, state of being i.e. intoxication level), characteristics of the vehicles involved in the accident and type of injuries acquired in the accident. Data processing through logistic regression models was preceded by a detailed statistical analysis that brought to light the incidence of factors in play in the series of accidents that are the object of study.

Accident data differentiated by road type demonstrate a strong localization of accidents on freeway (35.83%), built-up areas (32.40%) and state roads (16.78%).

Analyzing the distribution of accidents by road characteristics it becomes evident that more than half of the total number of accidents (60.65%) occurs on roads with straight stretches. Roads with curves account for 21.97% of the total.

Statistical analysis clearly demonstrates that the accident risk factor does not depend on the visibility factor. In fact, 85% of accidents occur in situations of good visibility.

Figure 4 illustrates road situations at the moment of an accident in relation to weather conditions: 63% of accidents occur in clear weather conditions and the road surface is dry in the majority of cases (70%). Difficult weather conditions such as fog/mist and snow occur very rarely whereas rain is present in only 19% of the cases and the percentage of accidents that occur on wet road surfaces account for 22% of the total.

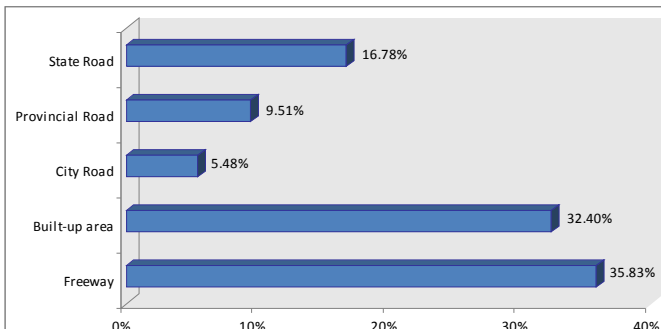


Figure 2: Incident distribution by road type.

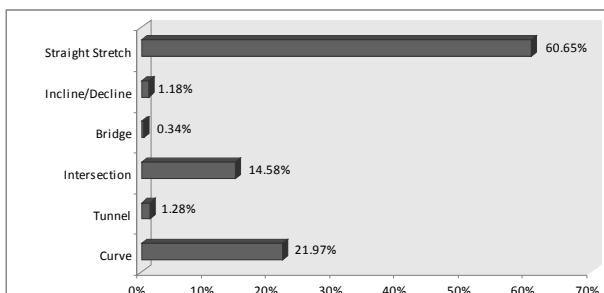


Figure 3: Incident distribution by road characteristics.

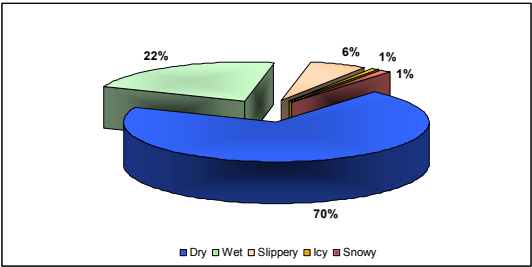


Figure 4: Incident distribution by road surface.

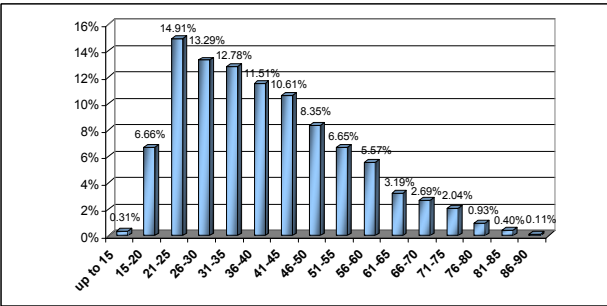


Figure 5: Incident distribution by age.

Analyzing the distribution of accidents by type, it can be shown that 40% of accidents occurred in the province of Cosenza over the three year period 2004 – 2005 – 2006 and the most common accident type was the “front – side” impact. Pile-ups were also among higher accident types (18%) as well as crashes against stationary objects (12%).

The age span of drivers involved in road accidents stretches from 21 – 25 years of ages (13.3%), 26 – 30 years of age (13.2%), 31 – 35 years of age (12%), 36 – 40 years of age (11%) and 41 – 45 years of age (10%).

4 Proposed logistic regression models and results

The data used was processed to develop logistic regression models that can be valid evaluation tools for the reliability of a road way.

Two models were specified, calibrated and validated:

- a logistic regression model to consider the A3 freeway in the province of Cosenza from the north side of the Laino Borgo municipality to the south side of Grimaldi;
- a logistic regression model for state roads in the province of Cosenza.

In the logistic regression model, the dependent variable Y (reliability) is considered dichotomous and functions together with independent variables X .



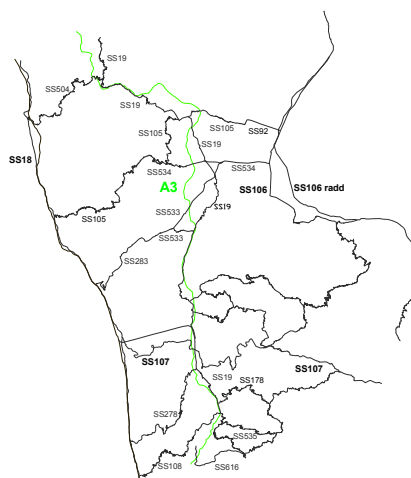


Figure 6: Roadways under investigation.

The dichotomous expression excludes, or minimizes, subjective elements so that the model fixed the reference conditions, allows an unambiguous definition if the conditions, state, behaviour etc., are or are not congruent with the pre-fixed reference conditions. In this way they are considered as respectful or disrespectful of technical norms, in the road project phase; appropriate or inappropriate behaviour of the road user, regarding road signage, the rules and road code, the environmental and traffic conditions; whether the overtaking distance or braking distance exists or not (in relation to visibility); whether there is fog, rain or snow present; proper or poor maintenance of road pavement etc.

The value of Y in logistic regression represents the probability that a variable response can take on a harder condition than others in relation to specific values of explicative variables x : $P[Y=1|x]$ or $P[Y=0|x]$, this estimation ranges between 0 and 1.

$$\text{logit } \pi(x) = \beta_0 + \sum_i^q \beta_i x_i \quad (1)$$

where

$$\text{logit } \pi(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] \quad (2)$$

and $p(x)$ shows the probability that Y is worth 1 depending on the explicative variables x : $p(x) = P(Y=1|x)$.

The logit transformation is the functional tie that allows a description of the relationship between the dependent variables and the explicative variables. The probability of Y can be expressed as a logistic function in the form:

$$\pi(x) = \frac{e^{\beta_0 + \sum x_i \beta_i}}{1 + e^{\beta_0 + \sum x_i \beta_i}} \quad (3)$$

From the estimation of the parameters β of the observed data set, the expected value of $E(Y_0|x_0)$ is expressed by the formula:

$$E(Y_0/x_0) = \frac{e^{\beta_0 + \sum x_{0i} \beta_i}}{1 + e^{\beta_0 + \sum x_{0i} \beta_i}} \quad (4)$$

In the definition of the logistic regression model, six explicative variables of the system components have been considered that are believed to have a greater affect on the reliability of the mobility system and road safety.

The choice of the variables is coherent with the set objective because the research concerns the global reliability of the path and not the single components. The variables are: structural (1), maintenance (2), behaviour (3), environment (4), age (5), gender (6). For each variable it is necessary to provide the following specifications:

- the structural variable (1) contains geometric characteristics: straight stretch roads, curves, tunnels, intersections, bridges, exits and inclines/declines;
- the maintenance variable (2) consists of two conditions: optimal maintenance and non-optimal maintenance of road pavement;
- the behaviour variable (3) consists of two conditions: driver behaviour according and not according to the law;
- the environment variable (4) considers road pavement dry or wet;
- the age variable (5) concerns the age of the driver;
- the gender variable (6) considers if the driver is male or female.

The definition of the system's components and of the corresponding explicative variables ($x_1, x_2, x_3, \dots, x_n$) has allowed the general formulation of a logistic regression model according to:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6}} \quad (5)$$

where it has been indicated with:

β_0 : regression constant,

β_1 : regression coefficient related to the structural variable containing geometric conditions,

β_2 : regression coefficient related to the maintenance variable,

β_3 : regression coefficient related to the behavioural variable,

β_4 : regression coefficient related to the environment variable,

β_5 : regression coefficient related to the age variable,

β_6 : regression coefficient related to the gender variable.

In the parameters estimation of logistic regression, the maximum likelihood algorithm (maximum likelihood – ML) is used as it values the model parameters

in a way that maximizes the function (log – likelihood function) that indicates the probability to obtain the expected value of Y given the value of the independent variables.

In the application of the model, the dependent variable Y was associated with the six indicated variables and the set of variables (independent and dependent) were treated as dichotomous variables with the assumption of values of 1 and 0, with the following explanations:

- to the structural variable containing geometric conditions, a value of 1 has been assigned when there are straight stretch roads and a value of 0 when there are curves, tunnels, bridges, inclines and declines;
- to the maintenance variable, a value of 1 has been given when there is a 'normal' state of maintenance and a value of 0 when the state of maintenance is in 'disrepair';
- to the behavioural variable, a value of 1 has been assigned when the behavioural condition of the driver is "not in accordance with the law" and a value of 0 when the behavioural condition of the driver is "in accordance with the law";
- to the environment variable, a value of 1 has been assigned when there are dry road surfaces and a value of 0 when there are wet road surfaces;
- to the age variable, a value of 1 has been assigned if the driver's age is less than 50 and a value of 0 when the driver's age is over 50;
- to the gender variable, a value of 1 has been assigned if the driver is male and a value of 0 if the driver is female.

The assigning of values 1 or 0 to the structural variable containing geometric conditions, maintenance and environment originated from the information analysis obtained from the accident files from the province of Cosenza.

Regarding the behaviour variable, values 1 or 0 originated from the information analysis of the accident descriptions with the indications of probable causes determined by the accident, including the driver's behaviour.

Concerning the dependent variable Y , by the end of the application of the model, it was considered as a dependent "observed" variable, assigning the value $Y=1$ when there are road accidents with injury or death and a value of $Y=0$ when there are road accidents with only minor damages.

From data processing, through regression analysis, the coefficients β of regression for single independent variables have been generated. The results reported from the calibration of the models regarding the A3 freeway and the state roads in the province of Cosenza follow:

It can be seen in the table that, for both models, in the moment in which the explicative variables have favourable values, reliability decreases. This result confirms the indications provided from the statistical analysis and explains, keeping in mind that when all the conditions are negative, drivers instinctively adopt an increased sense of attentiveness and better behaviour.

In the moment in which the varying conditions are optimal, the driver experiences an excessive level of trust in the safety level, followed with an increase in velocity. In these conditions, should an accident occur, negative

Table 1: Calibration results of logistic regression models.

| <i>Coefficient</i> | <i>Model 1 (Freeway)</i> | | <i>Model 2 (State Road)</i> | |
|------------------------|--------------------------|------------------|-----------------------------|------------------|
| | <i>Value</i> | <i>t-Student</i> | <i>Value</i> | <i>t-Student</i> |
| β_0 | -3.359 | 137.364 | -5.134 | 236.190 |
| β_1 | 0.433 | 12.702 | 0.915 | 50.466 |
| β_2 | 2.901 | 321.500 | 2.104 | 67.880 |
| β_3 | 0.797 | 40.270 | 0.515 | 16.201 |
| β_4 | 0.864 | 46.129 | 1.033 | 75.203 |
| β_5 | 1.285 | 100.957 | 1.308 | 83.637 |
| β_6 | 0.412 | 4.228 | 0.631 | 24.068 |
| <i>N. observations</i> | <i>2314</i> | | <i>1555</i> | |
| ρ^2 | <i>0.468</i> | | <i>0.448</i> | |

results are heightened. It is also important to consider that the database used is comprised only of the conditions (behaviour, geometric conditions, maintenance, environment and traffic) when there are accidents whereas the conditions in absence of accidents (in such a case, maximum reliability) are not dealt with.

The decrease in reliability and the occurrence of an accident, serious or minor, is often not exclusively due to one sole component of a system, but to causes connected to many components that come about with different affective levels.

5 Conclusions

The methodological approach used is valid for the reliability analysis of a roadway. The results obtained are congruent with the objectives formulated at the beginning of this work. Successive research developments will regard the refining of the model, broadening the database with the file values that will be provided by the Road Police.

The level of information detail concerning the accidents could be increased, making it possible to use as a tool to point out with precision how the accident occurred and also to identify the characteristics of a location in the time a fatality occurs.

The possession of punctual information surveyed with rigorous methods constitutes a fundamental presupposition for the localization of risk areas and for the adoption of strategic precautionary planning so as to improve system reliability and reduce accident risk. In any case, the results obtained demonstrate the interesting nature of a global approach to the problem of evaluating the reliability of a road system; these have otherwise confirmed the validity of the choice to use the logistic regression model. The parameters considered each possess their own value with respect to the evaluation of reliability.

The parameters considered (geometric characteristics, environment conditions, pavement conditions, driver behaviour, gender and age) are often

present contemporaneously in the unfavourable occurrence of an accident; these parameters entwine mutually creating the conditions for an accident. Therefore, it is hardly realistic to attribute an accident to a sole cause; the presence of rain and any other plausible weather condition is not generally the only cause of an accident as it is essential to examine erred driver behaviour and/or poor vehicle maintenance, etc. Often serious accidents occur even with but a single negative element.

In particular, appropriate information to a driver during travel by means of electronic road signs (i.e. Variable Message Panels), mobile phones, web sites and other mobility information tools, can adequately advise drivers about appropriate road behaviour according to the operative conditions of the system in specific moments and location. Therefore, drivers can take advantage of a heightened level of safety and avoid unnecessary risks.

References

- [1] Christiansen, C.L., Morris, C.N. & Pendleton, O.J., *A Hierarchical Poisson Model with Beta Adjustments for Traffic Accident Analysis* (Technical Report 103). Austin, TX: Center for Statistical Sciences, The University of Texas at Austin, 1992.
- [2] MacNab, Y.C., A Bayesian hierarchical model for accident and injury surveillance. *Accident Analysis and Prevention*, 35(1), 91-102, 2003.
- [3] Miaou, S.P. & Song, J.J., Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis and Prevention*, 37(4), 699-720, 2005.
- [4] Liu, J., Ravishanker, N., Ivan, J.N. & Qin, X., Hierarchical Bayesian estimation of safety performance functions for two-lane highways using Markov chain Monte Carlo modeling, *J. Transport. Eng.* **131** (5), pp. 345–351, 2005.
- [5] Pawlovich, M.D., Li, W., Carriquiry, A. & Welch, T.M., Iowa's experience with road diet measures: use of Bayesian approach to assess impacts on crash frequencies and crash rates, *Transport. Res. Record* **1953**, pp. 163–171, 2006.
- [6] Bijleveld, F.D., The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. *Accident Analysis and Prevention*, 37(4), 591-600, 2005.
- [7] Song, J.J., Ghosh, M., Miaou, S. & Mallick, B., Bayesian multivariate spatial models for roadway traffic crash mapping, *J. Multivariate Anal.* 97, pp. 246–273, 2006.
- [8] Ma, J. & Kockelman, K.M., Crash modeling using clustered data from Washington State: prediction of optimal speed limits, *Proceedings of the IEEE Intelligent Transportation Systems Conference* Toronto Canada, 2006.
- [9] Park, E.S. & Lord, D., Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. *Transportation Research Record*:



- Journal of the Transportation Research Board, No. 2019, TRB, National Research Council, Washington, DC, pp. 1–6, 2007.
- [10] Ma, J. & Kockelman, K., Bayesian multivariate Poisson regression for models of injury count, by severity, *Transport. Res. Record* **1950**, pp. 24–34, 2006.
- [11] Aguero-Valverde, J. & Jovanis, P.P., Bayesian Multivariate Poisson Log-Normal Models for Crash Severity Modeling and Site Ranking. Paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C., 2009.

