Abstract

While travel demand models have become more sophisticated, no amount of sophistication can eliminate the uncertainties inherent in forecasting ridership and revenue. Some of the uncertainty comes from the model estimation, while other elements come from unknown sources. Decision makers that rely on point forecasts—especially forecasts that represent an “average” or “most likely” outcome, are failing to recognize that uncertainty accompanies these predicted outcomes. This “most likely” outcome will be as likely to be below forecast as it is to be above. Furthermore, the documented tendency for actual ridership and revenue to be below forecasted ridership and revenue, points to the need for greater care in dealing with uncertainty. The approach taken in forecasting ridership and revenue for the California High Speed Rail project was to build the best behavioral state-of-the-practice model possible, while also recognizing and then quantifying the sources of uncertainty in the model estimation as well as in the underlying assumptions about future conditions in society. Our forecasts used a methodical approach to identify all potential risks, and narrowed the evaluation to those that were most important. We then developed regression models of forecasted ridership and revenue as dependent variables and the identified risk factors as independent variables and coupled them with a Monte Carlo simulation. This process generated a statistical evaluation of 5,000 combinations of factors through the leveraging of a relatively small number of full model runs. The result was ridership and revenue forecasts that are expressed in probabilistic terms. This approach provides a greater understanding of the risks inherent in forecasting to decision makers.
equations and building a Monte Carlo simulation represents a very innovative approach for developing and interpreting travel demand forecasts.

**Statement of Innovation**

The work described in this paper was performed by Cambridge Systematics a part of their work as ridership and revenue forecasting subconsultant to the California High Speed Rail Authority (CHSRA). Thus, the authors’ financial interests were the same as any consultant working for a public client.
Development of a Risk Analysis Model for Producing High Speed Rail Ridership and Revenue Forecasts

12/13/2013

Word Count = 2,227 (words) + 3 * 250 (tables and figures) = 2,977

Authors

Rachel Copperman (Corresponding Author)
Cambridge Systematics, Inc.
10415 Morado Circle, Building II, Suite 340
Austin, TX 78759
Phone: (512) 691-8508
Fax: (512) 691-3289
E-mail: rcopperman@camsys.com

Jeffrey Buxbaum
Cambridge Systematics, Inc.
100 CambridgePark Drive, Suite 400
Cambridge, MA 02140
Phone: (617) 234-0534
Fax: (617) 354-1542
E-mail: jbuxbaum@camsys.com

Weimin Huang
Cambridge Systematics, Inc.
38 East 32nd Street, 7th Floor
New York, NY 10016
Phone: (212) 209-6640
Fax: (212) 209-6641
E-mail: whuang@camsys.com

Kimon Proussaloglou
Cambridge Systematics, Inc.
115 South LaSalle Street, Suite 2200
Chicago, IL 60603
Phone: (312) 346-9907
Fax: (312) 346-9908
E-mail: kproussaloglou@camsys.com

Prepared for presentation
Transportation Research Board
5th Innovations in Travel Demand Forecasting Conference
Baltimore, Maryland
April 2014
ABSTRACT

While travel demand models have become more sophisticated, no amount of sophistication can
eliminate the uncertainties inherent in forecasting ridership and revenue. Some of the
uncertainty comes from the model estimation, while other elements come from unknown
sources. Decision makers that rely on point forecasts—especially forecasts that represent an
"average" or "most likely" outcome, are failing to recognize that uncertainty accompanies these
predicted outcomes. This "most likely" outcome will be as likely to be below forecast as it is to
be above. Furthermore, the documented tendency for actual ridership and revenue to be below
forecasts of ridership and revenue, points to the need for greater care in dealing with
uncertainty.

The approach taken in forecasting ridership and revenue for the California High Speed Rail
project was to build the best behavioral state-of-the-practice model possible, while also
recognizing and then quantifying the sources of uncertainty in the model estimation as well as
in the underlying assumptions about future conditions in society. Our forecasts used a
methodical approach to identify all potential risks, and narrowed the evaluation to those that
were most important. We then developed regression models of forecasted ridership and
revenue as dependent variables and the identified risk factors as independent variables and
coupled them with a Monte Carlo simulation. This process generated a statistical evaluation of
5,000 combinations of factors through the leveraging of a relatively small number of full model
runs. The result was ridership and revenue forecasts that are expressed in probabilistic terms.
This approach provides a greater understanding of the risks inherent in forecasting to decision
makers.

STATEMENT OF INNOVATION

The approach taken in forecasting ridership and revenue for the California High Speed Rail
project was to build the best behavioral state-of-the-practice model possible, while also
recognizing and then quantifying the sources of uncertainty in the model estimation as well as
in the underlying assumptions about future conditions in society. While risk analysis models
have become common for assessing traffic and revenue forecasts, they are rare within transit
forecasting. Many of the risk factors associated with long-distance travel are different from
those connected with toll roads. Thus, the full risk analysis process from choosing risk factors
and assessing their underlying distribution to estimating regression equations and building a
Monte Carlo simulation represents a very innovative approach for developing and interpreting
travel demand forecasts.

STATEMENT OF FINANCIAL INTEREST

The work described in this paper was performed by Cambridge Systematics a part of their work
as ridership and revenue forecasting subconsultant to the California High Speed Rail Authority
(CHSRA). Thus, the authors’ financial interests were the same as any consultant working for a
public client.
INTRODUCTION AND MOTIVATION

Travel demand models and processes used to forecast ridership and revenue on new transit lines or toll roads have become more sophisticated, but this sophistication does not eliminate the inherent uncertainty at all stages of the forecasting process. Uncertainty stems from model estimation as well as from assumptions that underlie forecasts.

Travel demand models are based on cross-sectional snapshots of travel patterns and traveler behavior. Such models assume that travelers’ responses to future travel options will be the same as when the snapshot was taken—a reasonable assumption for short-term forecasts. In the case of long term forecasts, this approach ignores the likelihood of significant changes in traveler behavior resulting from structural changes in society and the economy. While we cannot know for certain what changes will occur in society, we can speculate.

Flyvbjerg who has studied inaccuracy in travel demand forecasting (2005) and published on large infrastructure projects, recently discussed quality control and due diligence in project management by taking an outside view to get decisions right (2013) (1,2). Other recent investigations include a study by Armstrong et al. (2013) who discuss the golden rule of forecasting and by Adler et al. (2013) who gives more detail on the risk associated with travel demand models and the importance of developing a comprehensive quantitative risk analysis model (3,4).

Recognizing these uncertainties, we incorporated risk analysis into the ridership and revenue forecasts for the planned high-speed rail system in California, predicated on the following concepts:

- The Version 2.0 Ridership and Revenue (R&R) model incorporated a Year 2012-2013 travel survey, a streamlined model structure, improvements in model estimation, calibration to new data, extensive validation to Year 2010 and Year 2000 conditions, and extensive sensitivity testing.

- The state-of-the-practice R&R model produces reasonable forecasts with reasonable sensitivities to changing conditions.

- Models are not perfect, and their imperfections need to be understood and reflected in the forecasts used for planning purposes.

- Future conditions cannot be known with certainty. The forecasts used for planning purposes need to recognize the uncertainties in the exogenous variables used as input to the model and present a reasonable range.

- Varying a number of factors directly within a sophisticated travel demand model, that can take hours or days to run, is infeasible given a finite budget and schedule.
METHODOLOGY AND RESULTS

In the risk analysis approach we used a range of assumptions for the factors that we believe have the greatest impact on high-speed rail ridership and revenue. We ran the R&R Model numerous times for each forecast year to provide a range of outcomes for development of regression models of both ridership and revenue. Working with assumptions regarding the probability distribution of the individual risk factors, we prepared a Monte Carlo simulation for each of the project alternatives.

Selecting Risk Factors

We compiled a comprehensive list of factors that may affect high-speed rail ridership and narrowed the list to those that we thought would have the greatest influence on ridership. We kept the number of risk factors to a reasonable level since each new risk factor significantly multiplies the amount of computation time and analysis. We eliminated risk factors that we believe are too speculative or difficult to address within the modeling framework, and selected risk factors that are largely independent of each other to reduce the complications caused by correlation.

We finalized the list to six factors not under the control of engineers and planners designing the HSR system that we believed had the greatest potential variability and/or influence on total high-speed rail ridership:

1. Total California population, households, and employment
2. Spatial distribution of population and employment.
3. Auto operating cost.
4. Airline fares.
5. High speed rail main mode choice constants (HSR constants).
6. Trip frequency model constants.

Range of Risk Factor Values and Distributions

We quantified each factor so that it could be treated as a continuous independent variable within a regression model. For each risk factor, we developed a low, middle, and high value for each forecast year, and then developed a probabilistic distribution around these values based on available research and analysis (see Table 1). Additional detail on the derivation of the distribution of the HSR constants is provided in the next section.
<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Risk Factor Quantitative Value</th>
<th>Level</th>
<th>Description</th>
<th>Regression Model Inputs</th>
<th>Distribution Description</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Population and Employment Growth</td>
<td>Ratio of future year households to observed year 2010 households</td>
<td>High</td>
<td>CSTDM Forecast - High household and employment growth rate</td>
<td>1.148 1.208 1.232 1.372</td>
<td>Correlated w/ Regional Spatial Distribution</td>
<td>Based on an analysis of historical county-level socioeconomic estimates and forecasts from many sources. Correlation between overall growth and regional spatial distribution is based on the general principal that any departure from “average” statewide socioeconomic growth will depend on the fortunes of the San Joaquin Valley.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>Mid-level household and employment growth rate</td>
<td>1.132 1.183 1.199 1.305</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>Low household and employment growth rate</td>
<td>1.098 1.131 1.141 1.191</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regional Spatial Distribution</td>
<td>Ratio of San Joaquin Valley population to rest of California</td>
<td>High</td>
<td>CSTDM Forecast - High growth rate in San Joaquin Valley</td>
<td>0.115 0.120 0.122 0.134</td>
<td>Correlated w/ Regional Spatial</td>
<td>Forecasts based on U.S. Energy Information Administration (EIA) projections for gasoline prices and fuel efficiency forecasts (5).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>Mid-level growth rate in San Joaquin Valley</td>
<td>0.103 0.105 0.106 0.112</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>Low growth rate in San Joaquin Valley</td>
<td>0.101 0.102 0.103 0.107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Auto Operating Cost</td>
<td>$/mile (2005$)</td>
<td>High</td>
<td>Based on high fuel forecasts and low fuel efficiency</td>
<td>$0.26 $0.24 $0.24 $0.24</td>
<td>Triangular, with Low set to 15% and High at 85%</td>
<td>Based on airline competitive response scenarios developed by Cambridge Systematics and Aviation Systems Consulting (6).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>Reference/Base</td>
<td>$0.21 $0.20 $0.19 $0.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>Based on low fuel forecasts and high fuel efficiency</td>
<td>$0.18 $0.17 $0.16 $0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Airline Fares</td>
<td>Air Fare Skim Factor</td>
<td>High</td>
<td>16% increase from base scenario</td>
<td>1.16 1.16 1.16 1.16</td>
<td>Triangular, with Low set to 15% and High at 85%</td>
<td>Based on analysis of long-distance trip rates (7).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>Base scenario</td>
<td>1.00 1.00 1.00 1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>9% reduction from base scenario</td>
<td>0.91 0.91 0.91 0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Speed Rail Main Mode Choice Model Constants</td>
<td>Change in HSR constant units from Base</td>
<td>High</td>
<td>Equivalent to 60 fewer minutes of IVTT for business comercial (90 for recreation/other)</td>
<td>0.61 0.61 0.61 0.61</td>
<td>Normal distribution w/ Mean = 0 and STD = 0.48</td>
<td>0.5th percentile value of the distribution to corresponds to the CVR constant, with assumption that error in midpoint is symmetrically distributed.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>Average of Offset Approach for CVR and Air Offset Method</td>
<td>0 0 0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>Equivalent to 60 more minutes of IVTT for business/commercial (90 for recreation/other)</td>
<td>-0.61 -0.61 -0.61 -0.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trip Frequency Model Constants</td>
<td>Annual average roundtrips per capita</td>
<td>High</td>
<td>Increase from Mid scenario of 1.75 round trips per person</td>
<td>9.11 9.11 9.11 9.11</td>
<td>Normal distribution w/ Mean = 7.36 and STD = 0.85</td>
<td>Based on analysis of long-distance trip rates (7).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mid</td>
<td>Constants calibrated to CHTS trip rates that produce average of 7.36 round trips per person</td>
<td>7.36 7.36 7.36 7.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>Decrease from Mid scenario of 1.75 round trips per person</td>
<td>5.61 5.61 5.61 5.61</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
High-Speed Rail Main Mode Choice Constants Distribution

An important part of any mode choice model is a modal constant that explains factors that are not quantifiable by the stated and revealed preference (RP) surveys. When dealing with existing modes such as auto, conventional rail (CVR), and air, we can calibrate this constant by comparing the model outcomes to observed behavior. With a new mode like HSR, this is impossible, so there is uncertainty in the specified constant.

Uncertainty in the HSR constants comes from the distributional assumptions of the model itself and the data used to estimate the model. The former is relatively straightforward, in that the logit model may not be an accurate representation of how individuals actually make mode choices. The latter refers to the uncertainties associated with how the stated preference (SP) data were collected, the survey instrument, respondents perceptions based on “public opinion” at time of the survey, and other related issues.

Mid-level HSR constants were specified based on the relationships of the air, CVR, and HSR constants estimated using SP data, and the air and CVR constants after calibration to match observed 2010 travel (8). The normal distribution was chosen to represent the uncertainty in the HSR constants. In developing a variance for the HSR constant distribution, we assumed that none of the unobserved characteristics of the HSR mode would make it worse than the CVR mode; thus, the CVR constant would represent the minimum value of the constant for HSR. We chose the 0.5th percentile value of the distribution to correspond to the CVR constant for the recreation/other purpose. In other words, we assumed that there was a 1 in 200 chance that the HSR constant would be less than the CVR constant.

Model Runs

Once the risk factors and their distributions were defined, the full ridership and revenue model was run to develop the input data for the estimation of the risk analysis regression equations. To limit the number of model runs to a reasonable amount, we used a fractional 2-level factorial design for running the full model (where the two levels correspond to High and Low from Table 1). Thirty-two runs for each forecast year were used to estimate all the main effects and two-factor interactions. An additional fifteen runs with data points surrounding the “mid-level” and between the “mid-level” and the “low” and “high” values for each risk factor were added to provide information regarding the non-linearity of the forecast distributions and to ensure that the regression models represented the middle values within the distributions, and not just the extremes. These additional runs were important since non-linear transformations were used to develop the regression models.

Risk Analysis Regression Models for Ridership and Revenue

Regression Models

Using the ridership and revenue forecasts from each of the travel demand model runs for each forecast year, we estimated relationships between levels of the risk factor inputs and the resulting revenue forecasts. The Monte Carlo method makes it feasible to estimate a probability distribution of possible outcomes with using a limited number of full model runs using a deterministic equation (in our case, the regression model). We analyzed both linear and non-
linear transformations of model variables, and found that an exponential relationship between revenue and risk factors resulted in the best model fit, with all forecast years having R-square values above 0.99. The differences between predicted revenues versus estimated revenues from the full model runs were between ±5 percent. For each of the years, the following functional form was used for the regression model:

\[
Revenue = \exp(\text{Intercept} + a \cdot \text{Overall Growth} + b \cdot \text{Regional Spatial Distribution} + c \cdot \text{Auto operating cost} + d \cdot \text{Airline fares} + e \cdot \text{HSR Mode Choice Constant} + f \cdot \text{Trip Frequency Constant})
\]

The coefficients and related statistical measures for the Year 2029 data are shown in Table 2. The standardized estimates show the estimated change in revenue in standard deviation units when the corresponding independent variable is increased by one standard deviation. For each of the forecast years, the HSR mode choice constant has the highest standardized estimate, followed by the annual round trips/person and auto operating cost. The HSR constant contributes most of the variance for Year 2029, compared to the other risk factors.

**Table 2: Regression Equation for Year 2029**

| Parameter                        | Estimate | Standard Error | t Value | Pr > |t| | Standardized Estimate |
|---------------------------------|----------|----------------|---------|------|--|-----------------------|
| Intercept                       | XXX      | XXX            | 133.54  | <.0001|   | 0.000                 |
| Growth in Households            | 1.302    | 0.095          | 13.68   | <.0001|   | 0.107                 |
| Regional Spatial Distribution   | 0.876    | 0.413          | 2.12    | 0.0400|   | 0.016                 |
| Auto Operating Cost             | 1.631    | 0.103          | 15.87   | <.0001|   | 0.124                 |
| Airline Fares                   | 0.093    | 0.034          | 2.74    | 0.0091|   | 0.021                 |
| HSR Mode Choice Constant        | 0.791    | 0.007          | 116.32  | <.0001|   | 0.891                 |
| Annual Round Trips/Person       | 0.136    | 0.002          | 56.56   | <.0001|   | 0.433                 |
| Adjusted R-square               | 0.997    |                |         |       |   |                       |

1 As of this writing, the California High Speed Rail Authority 2014 business plan has not yet been published, and so we are excluding the ridership and revenue forecasts from this research brief. The intercept values have been crossed out so that revenue can not be calculated from the regression coefficients.

**Ridership vs. Revenue**

We began the analysis by analyzing the relationship between ridership and revenue through linear regression model. These regression models had r-squared values of 0.99. Since revenue and ridership were highly correlated, we developed regression equations for revenue only and used average HSR fare information derived from the full model runs to calculate the corresponding ridership for the risk analysis.

**Monte Carlo Simulation**

We conducted the Monte Carlo simulation using the Crystal Ball add-on to Excel that gives the user capability to run a randomized series of scenarios. We defined the scenarios by providing
distributions of the six risk factor values. The simulation software used these distributions to construct 5,000 unique scenarios where revenue was forecast using the regression equations described earlier.

**Range of Ridership and Revenue Forecasts**

The result of the above-described process was 5,000 forecasts of ridership and revenue for each analysis scenario, as shown in Figure 1. For Year 2029, high-speed rail ridership and revenue at the 5th percentile (i.e. 5% likelihood that ridership/revenue will be less than that value) is 50% lower than the 50th percentile value, while the 95th percentile is 90% higher than the 50th percentile value.

![Cumulative Distribution of Year 2029 Forecasted California High Speed Rail Revenue](image)

As of this writing, the California High Speed Rail Authority 2014 business plan has not yet been published, and so we are excluding the ridership and revenue forecasts from this research brief.

**Figure 1** Cumulative Distribution of Year 2029 Forecasted California High Speed Rail Revenue

**CONCLUSION AND IMPLICATIONS**

This paper presents the results of a risk-analysis approach to preparing ridership and revenue forecasts for the planned high-speed rail ridership system in California. This approach recognizes that even state-of-the-practice travel demand models that have been fully calibrated and validated are subject to uncertainty regarding the estimated variables within the models, the underlying data used for estimation, and the observed data collected for calibration of the existing model.

The large sensitivity of the HSR constant to overall revenue, which is driving the large probabilistic range in overall high-speed rail ridership and revenue, is a reflection of the overall
uncertainty of the attractiveness of HSR within the California transportation environment. This uncertainty is driven by the following:

1. HSR currently does not exist in California, and thus we are unable to calibrate the HSR constant to observed mode shares.

2. HSR does not exist in the United States with Amtrak’s Acela providing the closest fast rail option. Since Americans have little experience with HSR, we can not use observed data or experiences from other parts of the country to guide our knowledge in assessing Californians’ willingness to use HSR. Although we have gained insights from SP surveys on the attractiveness of HSR between destinations within California, there is inherent uncertainty due to the lack of actual HSR experience.

3. Uncertainty exists in the HSR system itself. The HSR constant captures all unobserved attributes and variables that affect an individual’s decision to use HSR that are not captured by other variables within the model. This includes wait and terminal time, the existence vs. non-existence of security checkpoints, attractiveness of the HSR stations and trains including food options, etc. While the SP survey describes the proposed amenities and out-of-vehicle travel times associated with HSR, the future system may not match these proposed characteristics.

4. Uncertainty exists in the mode choice model and the methodology used to calculate the HSR constant. Inherent uncertainty exists in all parts of model estimation including, but not limited to, the estimated variables within the mode choice model, the data resulting from the SP survey, the observed data collected for calibration of the existing model, and the methodology used for calculating the mid-level HSR constant.

REFERENCES


