How Data Imputation affects Crash Modeling Results

Supplemental Material

Modeling process:

MCMC estimations arrive at a stable estimate using a random process. We used 100,000 iterations with burn-in samples of 20,000 and used the Gelman-Rubin (G-R) statistic as an indicator of convergence.

Model results:

The following tables show the results of models including all data plus those with imputed AADT for 10%, 30%, 50%, and 70% missing data. Models for total crashes and for fatal and incapacitating injury crashes only were estimated. The residuals from the OLS estimates of AADT were included in the models with imputed data to control for error in the imputed estimates. When 70% of the data is imputed, this has a large effect on the estimates, but little effect in the other models.

Table S1. Base Model estimates with all data.

	M1: TOTAL CRASHES				M2: FATAL & INCAPACITATING INJURY CRASHES			
	Mean Coeff.	t-stat	credible interval		Mean Coeff.	t-stat	credible interval	
			2.5th %	97.5th %			2.5th %	97.5th %
Total width (ln)	-0.139	-2.396	-0.648	0.377	-0.504	-6.545	-1.128	0.100
Lane count (ln)	1.293	13.71	0.457	2.128	0.243	1.503	-1.019	1.511
Median width (ln)	-0.100	-6.651	-0.233	0.032	0.023	0.615	-0.298	0.345
Shoulder width (ln)	-0.058	-1.428	-0.429	0.288	0.160	1.757	-0.528	0.911
VKT (ln)	0.683	64.95	0.592	0.778	0.747	33.14	0.565	0.955
Sinuosity (ln)	-7.803	-2.619	-31.54	15.97	5.885	10.40	1.417	10.28
Constant	0.694	0.335	-15.82	17.18	-11.62	-30.04	-14.45	-8.630
Spatial Correlation (phi)	-0.038	-4.849	-0.108	0.026	-0.031	-4.766	-0.099	0.013
Observations	8071				8071			
Df	8062				8062			
Log likelihood	-23311.8				-2448.51			

Table S2. Model estimates with 10% imputation.

	M3: TOTAL CRASHES				M4: FATAL & INCAPACITATING INJURY CRASHES			
	Mean Coeff.	t-stat	credible interval		Mean Coeff.	t-stat	credible interval	
			2.5th %	97.5th %			2.5th %	97.5th %
Total width (ln)	-0.144	-2.664	-0.741	0.263	-0.382	-4.083	-1.108	0.323
Lane count (ln)	1.114	13.06	0.130	1.753	0.314	1.785	-1.075	1.686
Median width (ln)	-0.093	-6.378	-0.264	0.015	0.120	2.745	-0.234	0.494
Shoulder width (ln)	-0.171	-4.597	-0.638	0.098	0.243	2.236	-0.590	1.261
VKT (ln)	0.655	63.41	0.539	0.733	0.709	28.343	0.502	0.925
Sinuosity (ln)	-4.128	-3.594	-14.69	4.077	0.798	1.390	-3.585	5.364
Residuals(AADT)	0.000	18.48	0.000	0.000	0.000	2.671	0.000	0.000
Constant	-1.128	-1.403	-8.468	4.723	-9.060	-21.505	-12.628	-6.115
Spatial Correlation (phi)	-0.006	-1.333	-0.074	0.027	-0.023	-3.246	-0.091	0.024
Observations	8071				8071			
Df	8061				8061			
Log likelihood	-23291.34				-2476.85			

Table S3. Model estimates with 30% imputation.

	M5: TOTAL CRASHES				M6: FATAL & INCAPACITATING INJURY CRASHES			
	Mean Coeff.	t-stat	credible interval		Mean Coeff.	t-stat	credible interval	
			2.5th %	97.5th %			2.5th %	97.5th %
Total width (ln)	-0.095	-1.631	-0.601	0.414	-0.204	-1.973	-1.067	0.628
Lane count (ln)	1.243	13.98	0.458	2.015	0.394	2.274	-1.108	1.882
Median width (ln)	-0.104	-7.198	-0.232	0.022	0.115	2.525	-0.257	0.520
Shoulder width (ln)	0.018	0.468	-0.331	0.349	0.075	0.800	-0.664	0.899
VKT (ln)	0.654	63.84	0.564	0.746	0.751	29.15	0.540	0.976
Sinuosity (ln)	-18.22	-13.17	-28.96	-6.909	-6.504	-9.574	-11.42	-0.626
Residuals(AADT)	1.20E-05	14.33	5.00E-06	1.90E-05	4.30E-06	2.316	-1.20E-05	2.10E-05
Constant	7.814	8.100	-0.160	15.23	-4.888	-9.824	-8.860	-1.205
Spatial Correlation (phi)	-0.006	-1.412	-0.050	0.028	-0.041	-5.536	-0.114	0.009
Observations	8071				8071			
Df	8061				8061			
Log likelihood	-23438.37				-2481.25			

Table S4. Model estimates with 50% imputation.

	M7: TOTAL CRASHES				M8: FATAL & INCAPACITATING INJURY CRASHES			
	Mean Coeff.	t-stat	credible interval		Mean Coeff.	t-stat	credible interval	
			2.5th %	97.5th %			2.5th %	97.5th %
Total width (ln)	-0.214	-3.746	-0.724	0.280	-0.140	-1.746	-0.846	0.492
Lane count (ln)	1.221	12.95	0.401	2.066	-0.195	-1.200	-1.531	1.261
Median width (ln)	-0.133	-8.529	-0.272	0.004	0.060	1.377	-0.300	0.447
Shoulder width (ln)	0.017	0.405	-0.376	0.381	0.137	1.392	-0.698	0.969
VKT (ln)	0.697	60.650	0.596	0.801	0.744	26.92	0.521	0.988
Sinuosity (ln)	-0.647	-0.515	-9.943	10.032	1.357	1.627	-5.064	7.307
Residuals(AADT)	0.000	11.54	0.000	0.000	0.000	1.701	0.000	0.000
Constant	-4.058	-4.617	-11.552	2.423	-9.426	-19.54	-13.168	-5.747
Spatial Correlation (phi)	-0.030	-4.297	-0.100	0.025	-0.018	-2.854	-0.093	0.021
Observations	8071				8071			
Df	8061				8061			
Log likelihood	-23717.89				-2484.26			

Table S5. Model estimates with 70% imputation.

	M9: TOTAL CRASHES				M10: FATAL & INCAPACITATING INJURY CRASHES			
	Mean Coeff.	t-stat	credible interval		Mean Coeff.	t-stat	credible interval	
			2.5th %	97.5th %			2.5th %	97.5th %
Total width (ln)	-0.049	-0.878	-0.548	0.446	0.061	0.765	-0.598	0.667
Lane count (ln)	1.152	12.217	0.309	1.973	-0.313	-2.082	-1.535	0.849
Median width (ln)	-0.081	-5.212	-0.221	0.055	0.057	1.557	-0.243	0.380
Shoulder width (ln)	-0.080	-1.881	-0.470	0.284	0.213	2.459	-0.539	0.938
VKT (ln)	0.666	61.647	0.572	0.764	0.722	29.21	0.524	0.929
Sinuosity (ln)	-10.662	-8.068	-21.59	-0.520	11.60	19.96	7.355	16.362
Residuals(AADT)	39025.90	63.966	34372.39	43837.82	28560.14	54.23	24314.92	32736.18
Constant	2.622	2.863	-4.323	10.31	-17.17	-41.27	-20.30	-13.80
Spatial Correlation (phi)	-0.020	-3.358	-0.082	0.026	-0.009	-2.015	-0.059	0.025
Observations	8071				8071			
Df	8061				8061			
Log likelihood	-23812.88				-2501.55			

Impacts of variations in proportions of missing data:

Tables S1 – S4 show the results of varying the proportion of data for which AADT was imputed. Specific variable changes in direction of effect are discussed below.

Total Crashes:

For the total crashes models with 10% imputed AADT (Model 3), there is no change in the direction of effect for any variables when compared with Model 1 (base total crashes model). Model 5 (30% imputed AADT) showed a change in direction of effect for shoulder width.

The 50% model (Model 7) is consistent with the 30% model (Model 5), showing a change in direction for shoulder width from Model 1.

The 70% model (Model 9), unlike the 30% model (Model 5), there is no change in direction of any variable when compared with Model 1.

Fatal and Incapacitating Injury Crashes:

While Model 6 (30% imputed AADT) shows change in direction from the base model (Model 2) for sinuosity, Model 4 (10% imputed AADT) does not show any variable changes from Model 2.

Model 8 (50% imputed AADT) shows a change in direction for lane count only and Model 10 (70% imputed AADT) show a change in direction for both total width and lane count.