MARITIME TRANSPORTATION RESEARCH AND EDUCATION CENTER
TIER 1 UNIVERSITY TRANSPORTATION CENTER
U.S. DEPARTMENT OF TRANSPORTATION

Learning from USACE Open Data for Locks

9/13/2018 – 5/31/21

Justin R Chimka; University of Arkansas, Fayetteville

6/14/21

FINAL RESEARCH REPORT
Prepared for:
Maritime Transportation Research and Education Center

University of Arkansas
4190 Bell Engineering Center
Fayetteville, AR 72701
479-575-6021

ACKNOWLEDGEMENT
This material is based upon work supported by the U.S. Department of Transportation under Grant Award Number 69A3551747130. The work was conducted through the Maritime Transportation Research and Education Center at the University of Arkansas.

THANK YOU
Undergraduate Research Assistant: Benjamin Baser

1. Project Description

Note: Open Data for Locks were originally supposed to be made available via a Connected Government Cloud product sold by Tyler Technologies; USACE was a customer at some point, but that is no longer the case. Furthermore, Navigation and Civil Works Decision Support Center (NDC) was not actively pursuing public data dissemination as of Feb 2019. Therefore our focus became to analyze the NDC Key Lock Report, "a monthly summary and year-to-year totals of commodity tonnages and barge traffic for key locks on the inland waterways (USACE Institute for Water Resources Planning Assistance Library)."

This project had three objectives:

1. **Create a dataset to substitute for Open Data**; this objective will be described in 1. Project Description.
2. **Evaluate model selection strategies with a focus on the problem of interdependence among regressors**; this will be described in 2. Methodological Approach.
3. **Interpret the selected model(s) in order to learn from data, and recommend a way to choose Key Locks for priority preventive maintenance**; described in 3. Results/Findings.

The following reports are among those generated by NDC, using the Lock Performance Monitoring System (LPMS): Key Lock Report, Public Lock Commodity Report. Public Lock Usage Report, Public Lock Unavailability Report. Our analysis was purposefully limited to the 11 Key Locks. They are described along with metadata in **Appendix A**. We utilize Key Lock chambers, length and width as potential regressors. Public Lock Reports contribute the majority of our variables.

Public Lock Commodity Reports provide an annual summary of commodity movements from 1999 to the latest available year (2017). We utilize all nine (9) Public Lock Commodity variables as potential regressors.

- Commodities 10. Coal, Lignite, and Coal Coke
- Commodities 20. Petroleum and Petroleum Products
- Commodities 30. Chemicals and Related Products
- Commodities 40. Crude Materials, Inedible, Except Fuels
- Commodities 50. Primary manufactured Goods
- Commodities 60. Food and Farm Products
- Commodities 70. Manufactured Equipment and Machinery

- Commodities 80. Waste Material
- Commodities 90. Unknown or Not Elsewhere Classifies

Public Lock Usage Reports provide a summary of usage from 1999. We utilize the following 16 Public Lock Usage variables as potential regressors: Average Delay (Tows) (Hrs), Average Processing Time (Hrs), Barges Empty (#), Barges Loaded (#), Commercial Vessels (#), Commercial Flotillas (#), Commercial Lockages / Cuts (#), Non-Vessel Lockages (#), Non-Commercial Vessels (#), Non-Commercial Flotillas (#), Non-Commercial Lockages / Cuts (#), Percent Vessels Delayed (%), Recreational Vessels (#), Recreational Lockages (#), Total Vessels (#), Total Lockages (#).

Public Lock Unavailability Reports provide openings and closures by schedule type. We utilize Scheduled Unavailabilities (#) and Scheduled Unavailable Time as regressors, and Unscheduled Unavailabilities (#) as our response of interest.

In summary of the columns we are attempting to model Key Lock Unscheduled Unavailabilities with 30 regressors at our disposal: three (3) Key Lock Metadata, nine (9) Public Lock Commodity variables, 16 Public Lock Usage variables, and two (2) scheduled availability variables. As for rows we have 107 combinations of nine (9) Key Locks and 18 years (2000-2017). The remaining 55 combinations are a case of missing data as are 1999; Demopolis Lock and Dam, and Bonneville Lock and Dam. Therefore we have a dataset to substitute for Open Data. It is composed of (1 + 30) (107) = 3317 observations.

In the next section 2. Methodological Approach we evaluate model selection strategies with a focus on the problem of interdependence among regressors.

2. Methodological Approach

Interdependence among regressors – or multicollinearity – refers to the fact that variation in individual regressors can be explained by linear functions of other regressors that would participate a model of the response. When interdependence is not very serious regressors can peacefully coexist without masking each other's significance. Otherwise multicollinearity leads to variance inflation which artificially increases regressor standard errors, decreases test statistics, and increases p-values beyond the standard set to determine significance – even if the model itself is significant. Waiting for such a pathological result would seem like a liberal and reasonable approach especially if there was no physical reason to believe in multicollinearity. However, when

we know our regressors are mathematically or physically related a more proactive approach to mitigating interdependence may be appropriate. That is certainly the case in our data here. For example:

- In the Public Lock Usage data lockages / cuts, vessels and flotillas are physically related.
- In the Public Lock Usage data Total Vessels and Lockages are mathematically related to those numbers for Commercial, Non-Commercial and Recreational.
- In the Public Lock Unavailability data Scheduled Unavailabilities and Scheduled Unavailable Time are physically related.

In fact if we attempt best subsets regression with our dataset in Minitab, it returns an error associated with interdependence.

Instead of relying on our intuition and personal preferences to add and delete variables that contribute greatly to interdependence we test two data-driven systems for variable selection, and evaluate them based on an objective measure of model efficiency.

System A is a backward-selection algorithm based on regressor p-value or significance. In other words we begin with a full model of Unscheduled Unavailabilities, iteratively delete the regressor with greatest p-value, and refit until we have deleted all but one regressor.

System B is a backward-selection algorithm based on the Variance Inflation Factor (VIF). It quantifies the extent to which individual regressors can be explained by linear functions of other regressors that would participate in a model of the response. We begin with a full model of Unscheduled Unavailabilities, iteratively delete the regressor with greatest VIF, and refit until we have deleted all but one regressor. Where R-squared (j) is the coefficient of determination which results when x (j) is regressed versus all other regressors: VIF [x (j)] = 1 / [1 – R-squared (j)].

Our objective measure of model efficiency is <u>not</u> adjusted R-squared. Our longterm focus on modeling is to work with a reasonable number of variables versus sample size, and explore the interaction among regressors when some are insignificant (alpha = 0.05). Significant interaction involving generally insignificant regressors will cause us to divide an original dataset into more meaningful subsets that can eventually be described with only significant variables, while all deleted variables have been fully vetted for the possibility of significant interaction. **Our objective in model selection here is to minimize inefficiency (INE) = Insignificant regressors (#) – Significant**

regressors (#). In **Appendix B** are detailed results of Systems A and B with respect to INE and variables deleted at every iteration.

System A, traditional backward-selection based on p-values, produces a minimum INE = -13 at iteration 18 / 30. The model from System A has adjusted R-squared = 0.4360, zero (0) insignificant regressors and 13 significant regressors that will be described in detail next section. System B, backward-selection based on VIF, produces a minimum INE = -1 at iteration 7 / 30. The model from System B has adjusted R-squared = 0.2429, three (3) insignificant regressors and four (4) significant regressors that will be described in detail next section.

3. Results/Findings

Of all 13 significant regressors produced by System A three (3) of them appear in the model produced by System B: Commodities 10, Commodities 70, and Average Processing Time, so their directional effect and significance on Unscheduled Unavailabilities could be verified. All Terms and T-values for both models are reported in **Appendix C**.

- In the models produced by both Systems A and B Commodities 10 (Coal, Lignite, and Coal Coke) has a positive, significant effect on Unscheduled Unavailabilities.
- In the model produced by System A Commodities 70 has a positive, significant effect on Unscheduled Unavailabilities; in the model produced by System B Commodities 70 (Manufactured Equipment and Machinery) has a positive (but insignificant) effect on Unscheduled Unavailabilities.
- In the models produced by both Systems A and B Average Processing Time has a positive, significant effect on Unscheduled Unavailabilities.

It should be noted that the significance of these effects are controlling for all other variables in the model under consideration. For example the model produced by System B includes the variable Scheduled Unavailable Time which means Commodities 10 and Average Processing Time are significant for any amount of Scheduled Unavailable Time observed in the data used to fit the model. **For purposes of identifying Key Locks that should be prioritized for preventive maintenance we recommend looking for locks with combinations of large values in Commodities 10 and Average Processing Time**.

For purposes of predicting Unscheduled Unavailabilities we recommend the model produced by System A. All of its variables are significant, and it explains more than half

of the variation in Unscheduled Unavailabilities (R-squared = 0.5051): Unscheduled Unavailabilities ~ normal (m, s).

- E (m) = 250.1 – 99.2 (Chambers) – 2.087 (Width) + 9E-6 (Commodities 10) + 1.7E-5 (Commodities 30) + 7E-6 (Commodities 40) + 7E-6 (Commodities 60) + 2.7E-5 (Commodities 70) + 95.7 (Average Processing Time) – 8.92E-3 (Barges Loaded) + 3.12 (Non-Commercial Vessels) – 3.06 (Non-Commercial Lockages / Cuts) – 0.806 (Percent Vessels Delayed) + 0.1039 (Recreational Lockages)
- E (s) = 29.377

4. Impacts/Benefits of Implementation (actual, not anticipated)

Most research effort related to our focus is on vessel delay rather than lock unavailability (Zhang, et al., 2015; Yu, et al., 2019). Other related research deals with commodity flow and estimating economic value (Baroud, et al., 2014). There is also significant work showing the economic impact of major disruptions at the state and national (Folga, et al., 2009; Tong and Nachtmann, 2017) levels. Another main focus of work referenced here is the overall decline in infrastructure of the lock and dam system. It is known that many facilities are in need of upgrade or repair, but modernizing a lock means the system is temporarily unavailable, and there are various economic and usage factors that must be balanced for the potential downtime to be worthwhile (Dowd, et al., 2020).

Our primary contribution to the literature is a recommendation about how to prioritize Key Locks for upgrade and repair having identified consistently significant factors in models of Unscheduled Unavailabilities. The aforementioned models were systematically produced by variable selection algorithms implemented to mitigate interdependence among regressors. Our secondary contribution is an efficient equation which explains more than half of the variation in Key Lock Unscheduled Unavailabilities and could be used to predict them as a function of the following.

- Key Lock metadata: As Chambers and Width increase Unscheduled Unavailabilities tend to decrease.
- Commodities (10, 30, 40, 60, 70): As they increase Unschedule Unavailabilities tend to increase also.
- A variety of Usage variables have significantly positive or negative effects on Unscheduled Unavailabilities.

5. Recommendations and Conclusions

First a general observation: When in need of an iterative method to mitigate the effects of interdependence and select regressors for a multiple linear regression model it will not necessarily benefit from focusing on the measure of multicollinearity. We tried separate backward-selection algorithms based on the p-values and VIF of regressors in the models, and evaluated them according to a measure of efficiency. The system based on p-values produced a model 13x better with respect to efficiency as we have defined it, compared to the system based on VIF. On the scale of adjusted R-squared the system based on p-values produced a model which explains 19.31% more variation in Unscheduled Unavailabilities, compared to the system based on VIF. Finally the system based on p-values happened to converge 25% percent more quickly than the did the system based on VIF.

Key Locks should be prioritized for preventive maintenance according to their levels in Commodities 10 (Coal, Lignite, and Coal Coke) and Average Processing Time, with greater magnitudes of these measures being associated with significantly greater numbers of Unscheduled Unavailabilities, controlling for other variables including Scheduled Unavailable Time, in our best models of Unscheduled Unavailabilities. A secondary consideration in deciding how to prioritize Key Locks for preventive maintenance should be according to Commodities 70 (Manufactured Equipment and Machinery). Greater magnitudes of this measure are associated with greater numbers of Unscheduled Unavailabilities in our best models, and the relationship is significant in our best model.

To imagine how changes in parameters like Commodities 10 and 70, and Average Processing Time assuming they could be controlled, may affect future observations of Unscheduled Unavailabilities see our equation for E (m) at the end of Section 3.

In conclusion we would like to offer ideas for future work related to this project. Having identified Key Locks and integrated some of their metadata with relevant data from the Public Lock Reports, we are on the lookout for additional lock demographics and years of Commodities, Usage and Unavailability, especially as additional rows or years of data could build a sample size to better support analysis of spatial and temporal effects. Also future research into the regressions and systems for variable selection could include some model adequacy checking, especially if there is great interest in prediction intervals around estimates of Unscheduled Unavailabilities.

Finally on the subject of identifying Key Locks to prioritize for preventive maintenance one could use cluster k means to organize Key Locks into a number of groups based on average observations in Commodities 10 and Average Processing Time.

References

Baroud, H, K Barker, JE Ramirez-Marquez and CM Rocco S (2014), "Importance measures for inland waterway network resilience," Transportation Research Part E: Logistics and Transportation Review 62: (2014): 55-67.

Dowd, Z, AY Franz and JS Wasek (2020), "A decision-making framework for maintenance and modernization of transportation infrastructure," IEEE Transactions on Engineering Management 67(1): 42-53.

Folga, S, T Allison, Y Seda-Sanabria, E Matheu, T Milam, R Ryan and J Peerenboom (2009), "A systems-level methodology for the analysis of inland waterway infrastructure disruptions," Journal of Transportation Security 2 (121).

Tong, J and H Nachtmann (2017), “Cargo prioritization and terminal allocation problem for inland waterway disruptions,” Maritime Economics and Logistics 19: 403-427.

Yu, TE, BP Sharma and BC English (2019), “Investigating lock delay on the Upper Mississippi River: a spatial panel analysis,” Networks and Spatial Economics 19: 275-291.

Zhang, YY, M-S Chang and SW Fuller (2015), “Statistical analysis of vessel waiting time and lockage time on the Upper Mississippi River,” Maritime Economics and Logistics 17: 416-439.

Addendum

A review panel doubted two things about this research: 1) that commodity variation could physically affect unavailability, and 2) significant effects could be generalizable across key locks. These are interesting observations in that they seem to suggest commodity variables could be acting in part as surrogates for some unmeasured “lock demographics,” and by controlling for them we are better able to make general statements about the effects of key lock usage.

Appendix A. Key Locks and Metadata

Division	River	Lock	Chambers	Length	Width	Latitude	Longitude
South Atlantic	Blackwater / Tombigbee River	Demopolis Lock & Dam	1	600	110	32.520221	-87.880578
Northwestern	Columbia River	Bonneville Lock & Dam	1	500	76	45.6379405	-121.946994
Mississippi Valley	Gulf Intracoastal Waterway	Inner Harbor Navigational Canal	1	640	75	29.964756	-90.027228
Mississippi Valley	Gulf Intracoastal Waterway	Calcasieu Lock	1	1205	75	30.087041	-93.291994
Mississippi Valley	Illinois River	LaGrange Lock & Dam	1	600	110	39.940403	-90.534752
Great Lakes & Ohio River	Kanawha River	Winfield Locks & Dam Main 1	2	360	56	38.5272113	-81.9136757
Mississippi Valley	Mississippi River	Lock & Dam 25	1	600	110	39.003782	-90.689744
Mississippi Valley	Mississippi River	Chain of Rocks Lock & Dam 27	2	1200	110	38.701757	-90.1818375
Southwestern	McClellan-Kerr Arkansas River Navigation System	Norrell Lock & Dam	1	600	110	34.019341	-91.193476
Great Lakes & Ohio River	Ohio River	Lock & Dam 52	2	1200	110	37.126933	-88.655722
Great Lakes & Ohio River	Tennessee River	Kentucky Lock	1	600	110	37.015165	-88.265598

Appendix B. Results of Systems A and B for Variable Selection

Iteration	Adj R^2_A	INE_A	$Deleted_A$	Adj R^2_B	INE_B	$Deleted_B$
1	0.4043	18	Recreational Vessels	0.4043	18	Non-Commercial Lockages
2	0.4043	17	Total Vessels	0.4043	17	Total Lockages
3	0.4118	14	Commercial Lockages	0.3916	22	Commercial Flotillas
4	0.4118	15	Total Lockages	0.3970	13	Barges Loaded
5	0.4189	10	Average Delay	0.3804	18	Non-Commercial Vessels
6	0.4249	9	Scheduled Unavailable T.	0.3804	17	Total Vessels
7	0.4300	8	Length	0.3665	18	Commercial Lockages
8	0.4408	5	Commercial Flotillas	0.3737	17	Barges Empty
9	0.4456	4	Commercial Vessels	0.3675	14	Commercial Vessels
10	0.4456	1	Commodities 90	0.3695	15	Length
11	0.4466	-4	Commodities 20	0.3917	16	Commodities 40
12	0.4486	-3	Barges Empty	0.3726	13	Recreational Lockages
13	0.4497	-4	Commodities 50	0.3773	10	Commodities 20
14	0.4508	-7	Commodities 80	0.3839	9	Chambers
15	0.4498	-8	Scheduled Unavailabilities	0.3707	8	Commodities 30
16	0.4513	-9	Non-Vessel Lockages	0.3372	7	Width
17	0.4443	-10	Non-Commercial Flotillas	0.3371	6	Commodities 60
18	0.4360	-13	Commodities 70	0.3440	5	Commodities 90
19	0.4070	-12	Percent Vessels Delayed	0.3506	4	Percent Vessels Delayed
20	0.3803	-5	Average Processing Time	0.3548	3	Commodities 50
21	0.3737	-6	Commodities 40	0.2729	0	Scheduled Unavailabilities
22	0.3689	-7	Width	0.2387	1	Average Delay
23	0.3691	-8	Commodities 60	0.2353	0	Non-Commercial Flotillas
24	0.3463	-7	Non-Commercial Lockages	0.2429	-1	Commodities 80
25	0.3166	-2	Non-Commercial Vessels	0.1321	2	Commodities 10
26	0.3192	-3	Chambers	0.0398	5	Recreational Vessels
27	0.3060	-4	Barges Loaded	0.0140	2	Average Processing Time
28	0.0990	1	Commodities 30	0.0163	1	Scheduled Unavailable T.
29	0.1075	0	Recreational Lockages	0.0258	0	Non-Vessel Lockages
30	0.1008	-1	Commodities 10	0.0340	-1	Commodities 70

Appendix C. Models Produced by Systems A and B

Term	$T\text{-value}_A$	$T\text{-value}_B$
Chambers	-4.45	
Width	-3.26	
Commodities 10	5.98	4.47
Commodities 30	5.07	
Commodities 40	3.11	
Commodities 60	4.27	
Commodities 70	2.41	0.69
Commodities 80		3.96
Average Processing Time	2.95	2.67
Barges Loaded	-4.02	
Non-Vessel Lockages		-0.18
Non-Commercial Vessels	3.04	
Non-Commercial Lockages / Cuts	-2.94	
Percent Vessels Delayed	-2.59	
Recreational Vessels		2.84
Recreational Lockages	3.02	
Scheduled Unavailable Time		0.97