WORCESTER POLYTECHNIC INSTITUTE

Mathematical Cost Modelling

An Interactive Qualifying Project

Daniel Duhaney

2/29/2012

I use empirical data and multidimensional polynomial approximation to create a model that predicts tunnel outturn cost (O) given tunnel face area (f), length (l), depth (d), density of geology (r), with a certain error. I discuss the limitations of my methods and recommendations for improving outturn cost prediction models.

Table of Contents

List of Tables

List of Figures

Background:

This paper should be read in the context of the paper 'Analysing International Tunnel Costs' by Megan Read and Nathaniel Efron.

Introduction:

Many factors contribute to the cost of constructing a tunnel. Direct and quantifiable factors, such as tunnel length, face area, and depth affect the cost of constructing a tunnel. Indirect and often unquantifiable factors (e.g. market structure, contract types, and so on) also affect the cost of tunneling, though in ways less clear. Estimation of tunnel outturn cost (the final cost of constructing a tunnel) is difficult, especially in the feasibility assessment stage of tunnel delivery. Estimating the cost of a tunnel at this early stage relies on experience, familiarity and judgment, especially since very little detail about the tunnel has yet been developed. It is not surprising that estimation by these means is often unreliable. Even so, it is important to develop an estimate early on to aid in decision making.

Problem:

To create a model that provides a rough prediction of tunnel outturn costs given only basic information about a prospective tunnel (and thus can be used in an early stage).

Methods and Data:

Because of the generally limited understanding of the complex relationships between the direct and indirect factors that affect tunnel outturn cost, I decided to use a top-down approach to the problem. That is, I sought a model that predicts tunnel outturn cost through numerical analysis on empirical data on existing tunnels, rather than a model built on first principles (bottom-up).

I used a database of tunnels compiled by AECOM which stores basic data on tunnels from around the world. I reduced the information in the database into a collection of independent variables (inputs) and dependent variables (outputs). The inputs included tunnel dimensions (face area and length), depth, ground conditions (geology), location, end use, and setting (whether urban, rural, underwater, and so on). The outputs included tunnel cost estimates and outturn costs in terms of local currency and United States Dollars in the third quarter of the fiscal year of 2011 (3Q 2011 USD), adjusted for inflation. My method of numerical analysis required that I choose the fields that are or could be represented quantitatively. Therefore, I used the data on tunnel length, face area, depth, geology, and the tunnel outturn cost in 3Q 2011 USD. Unfortunately, I could not quantitatively characterize the 'location', 'end use', and 'setting' fields, and so they were excluded from my methods.

To quantitatively represent geology, I used the average of the density range [1] [2] of geological material (hereon refer to as substrate density). If ground conditions of a tunnel listed more than one geologic material, I averaged their densities. I discuss my reason for and the limitations of using the density over other geologic measures in the *Limitations and Recommendations* section.

The data set was reduced to the following table:

Face area (m^2)	Total tunnel length (m)	Depth (m)	Substrate Density (kg/m^3)	Tunnel Outturn cost cost per m^3 (USD)
3.45967891	12500	39.5	2067.5	527401421.3
8.042477193	2940	70	2450	106219588.5
9.23	6475	27.4	2050	285000000
9.621127502	3500	20	2365	364070520.3
11.04466167	1720	35	2265	58827057.12
12.56637061	2700	17.5	2615	160386268.6
13.20254313	3000	30	2450	113100795.5
30.9156876	6948	21	1520	1877777778
34.211944	5600	30	1806.7	1260316074
34.211944	5700	30	1806.7	978556.6729
34.211944	17000	30	1806.7	2043451963
40.71504079	25000	20	2365	2451439030
52.8	8500	53	2625	856550238.2
60.13204689	10400	17.5	1200	1233442623
68	600	10.75	1520	530000000
85.62410777	3400	65	1907.5	3375646567
86.29354105	9000	30	1.675	860812372.2
108.8247695	150000	45	1675	13251396648
109.3588403	9700	30	1550	727671294.8
128.6796351	9600	60	2300	2925333753
157.5	7200	35	2365	950285544
208.25	1220	15	1835	186650238.5
216	334	2	1200	569698449.6
264	2300	24	2615	909081273.9
284.75	1066	22.6	1333.3	232630894.3

Table 1 Dataset of tunnel outturn cost costs extracted from Tunnel Database

The face area *f,* length *l,* depth *d,* and density *r* were treated as independent variables. The tunnel outturn cost *O* was treated as a dependent variable. To establish a relationship between the independent variables and the outturn cost, I performed a multi-dimensional polynomial approximation of the data, using 25 tunnels, to obtain a polynomial function *O(f,l,d,r)* which represented the outturn cost in terms of the independent variables. I only used 25 of the 162 tunnels in the database because they were the only entries that had sufficient data for my methods. I did the approximation using the open-source computation software GNU Octave (version 3.6.0) with the MATLAB procedure `polyfitn` [3], developed by John D'Errico. This procedure will generate a polynomial of user-specified degree which best approximates the data. It also calculates the root mean squared error (RMSE) which is a measure of the difference between the model's predicted outturn cost and the actual outturn cost. I tried to fit polynomials of various degrees to determine the polynomial with a smallest RMSE. Finally, I tested the prediction power of the resulting polynomial by predicting outturn costs of other tunnels in the database. I searched the Internet for information to fill in missing data on some tunnels in order to test the model.

Results:

5-Dimensional Model

After trying to approximate the data with polynomials of degree 1 through 12, I determined with `polyfitn` that the third-degree polynomial below produced best fit the data:

$$O(f,l,d,r) = \alpha_1 f^3 + \alpha_2 f^2 l + \alpha_3 f^2 d + \alpha_4 f^2 r + \alpha_5 f^2 + \alpha_6 f l^2 + \alpha_7 f l d + \alpha_8 f l r + \alpha_9 f l + \alpha_{10} f d^2 + \alpha_{11} f d r + \alpha_{12} f d + \alpha_{13} f r^2 + \alpha_{14} f r + \alpha_{15} f + \alpha_{16} l^3 d + \alpha_{17} l^2 d + \alpha_{18} l^2 r + \alpha_{19} l^2 + \alpha_{20} l d^2 + \alpha_{21} l d r + \alpha_{22} l d + \alpha_{23} l r^2 + \alpha_{24} l r + \alpha_{25} l + \alpha_{26} d^3 + \alpha_{27} d^2 r + \alpha_{28} d^2 + \alpha_{29} d r^2 + \alpha_{30} d r + \alpha_{31} d + \alpha_{32} r^3 + \alpha_{33} r^2 + \alpha_{34} r - \alpha_{35}$$

The coefficients α_n are shown in the table below:

N	Coefficient α_n
1	-2.96255941113044E+04
2	-4.44377174701186E+01
3	-9.06755894474888E+05
4	3.44425203091521E+04
5	-2.39173190952942E+07
6	-3.08730662109830E-01
7	6.42937166533320E+02
8	-7.85808917956691E+01
9	-1.52320113557541E+04
10	-4.60783918046249E+06

n	Coefficient α_n
11	2.75032860951965E+05
12	7.66481875123373E+06
13	-3.50626095764755E+03
14	-1.84143646625789E+06
15	4.29347531122166E+09
16	-7.35181676226898E-03
17	1.37228425682256E+01
18	4.87248418192989E-01
19	4.71589798292065E+01
20	-3.69517689718862E+04

n	Coefficient α_n
21	1.24232966009511E+03
22	-3.17929064790973E+05
23	2.04339048921633E+00
24	-2.49370453702707E+04
25	-8.28926328654041E+06
26	7.75855830345857E+06
27	-1.35843214078291E+04
28	-7.55825606797954E+08
29	-9.21237973834501E+03
30	3.44696102332931E+07
31	-1.11607836988481E+10
32	1.46407991645032E+02
33	-5.28402897761441E+05
34	3.93425458953915E+08
35	9.43062442108226E+10

Table 2 Coefficients of *O(f,l,d,r)*

This polynomial had the least RMSE of the polynomials attempted, with an RMSE of $0.004836816 according to `polyfitn`.

Test Cases

I tested this polynomial by 'predicting' the outturn cost costs of the tunnels used in the dataset, as well outturn cost costs of other tunnels in the database not used in the dataset, chosen at random.

Data Fit

The results of the outturn cost cost-prediction test of tunnels used in the dataset:

Face area (m^2)	Total tunnel length (m)	Depth (m)	Substrate Density (kg/m^3)	Tunnel Outturn cost (USD)	*O(f,l,d,r)* Prediction of Outturn cost (USD)	Absolute Error

3.45967891	12500	39.5	2067.5	527401421.3	527401421.3	0.019579947
8.042477193	2940	70	2450	106219588.5	106219588.5	0.049663797
9.23	6475	27.4	2050	285000000	285000000	0.001510978
9.621127502	3500	20	2365	364070520.3	364070520.3	0.02457875
11.04466167	1720	35	2265	58827057.12	58827057.12	0.002805904
12.56637061	2700	17.5	2615	160386268.6	160386268.6	0.01638639
13.20254313	3000	30	2450	113100795.5	113100795.5	0.041209981
30.9156876	6948	21	1520	1877777778	1877777778	0.222022057
34.211944	5600	30	1806.7	1260316074	1260316074	0.432529449
34.211944	5700	30	1806.7	978556.6729	978556.6723	0.000566928
34.211944	17000	30	1806.7	2043451963	2043451963	0.468174934
40.71504079	25000	20	2365	2451439030	2451439030	0.211260319
52.8	8500	53	2625	856550238.2	856550238.2	0.028429031
60.13204689	10400	17.5	1200	1233442623	1233442623	0.048740387
68	600	10.75	1520	530000000	530000000	0.000656009
85.62410777	3400	65	1907.5	3375646567	3375646567	0.435409069
86.29354105	9000	30	1.675	860812372.2	860812372.2	0.001786828
108.8247695	150000	45	1675	13251396648	13251396648	0.052690506
109.3588403	9700	30	1550	727671294.8	727671294.8	0.01391387
128.6796351	9600	60	2300	2925333753	2925333753	0.349388123
157.5	7200	35	2365	950285544	950285544	0.043141842
208.25	1220	15	1835	186650238.5	186650238.5	0.038724899
216	334	2	1200	569698449.6	569698449.6	0.046922565
264	2300	24	2615	909081273.9	909081273.9	0.045301199
284.75	1066	22.6	1333.3	232630894.3	232630894.3	0.024805129
					Average error	0.104807956

Table 3 Comparison between actual and modeled outturn cost of tunnels in the dataset using the 5-D polynomial *O(f,l,d,r)*

We can see that the polynomial *O(f,l,d,r)* fits the data points in the dataset almost exactly. It predicts the outturn cost costs with a maximum error of $0.47 and with an average absolute error of $0.10.

Prediction Test

The results of the outturn cost-prediction test of tunnels not used in the dataset:

Tunnel	**Face area (m^2)**	**Total tunnel length (m)**	**Dept h (m)**	**Substrat e Density (kg/m^3)**	**Tunnel Outturn cost (USD)**	***O(f,l,d,r)* Prediction of Outturn cost (USD)**	**Absolute Error**
UK Tunnel 9	6.1575216	3300	21	1600	152124930.5	56558045749	56405920818
University Link (ULINK) Project	30.915688	6948	22.1	1573.3	1877777778	-2586634199	4464411977
Dulles Airport Train System	32.169909	1249	15	1935	1200000000	-75711934731	76911934731
UK Tunnel 1	52.16811	15000	21	1760	427663927.9	-1.33655E+11	1.34083E+11
Sants-Sagrera HSL	95.033178	5600	34	1985	256436718.3	32729662613	32473225895
Channel Tunnel	108.82477	150000	45	1825	13251396648	2.35023E+12	2.33697E+12
Bjørvika Immersed Tunnel	358.1	680	9	1200	857378542.6	-1.74647E+11	1.75505E+11
						Average Error	4.02402E+11

Table 4 Comparison between actual and modeled outturn cost of tunnels chosen at random using the 5-D polynomial *O(f,l,d,r)*

It is clear the predictions of *O(f,l,d,r)* in these cases are very inaccurate. The average absolute error was found to be $402,402,000,000. This is definitely not acceptable. Also, note that the model produces negative outturn costs, which is not a valuable prediction in the real-world.

This model, despite its low RMSE, still seems to be highly prone to error. I speculated that the complexity of the dataset and the model resulted in a highly-fluctuating polynomial. To overcome this I divided the outturn cost of each tunnel by its dimensions and simplified the dataset to compare depth and substrate density with outturn cost per meter-cubed *C*. This reduced the system from 5 to 3 dimensions, which allowed not only simpler analysis, but also allowed visualization of the stability of the resultant polynomial (which is very difficult with a 5-dimensional model.)

3-Dimensional Model

The simplified dataset:

Depth (m)	Substrate Density (kg/m^3)	Tunnel Outturn cost per m^3 (USD)
39.5	2067.5	12195.38426
70	2450	4492.286503
27.4	2050	4768.737163
20	2365	10811.6381
35	2265	3096.679501
17.5	2615	4727.086566
30	2450	2855.530544
21	1520	8741.892776
30	1806.7	6578.300307
30	1806.7	5.018031401
30	1806.7	3513.48221
20	2365	2408.386663
53	2625	1908.534399
17.5	1200	1972.330202
10.75	1520	12990.19608
65	1907.5	11595.3001
30	1.675	1108.377498
45	1675	811.7880213
30	1550	685.9770922
60	2300	2368.069087
35	2365	837.9943068
15	1835	734.6554561
2	1200	7896.685097
24	2615	1497.169423
22.6	1333.3	766.3840414

Table 5 Simplified dataset of tunnel outturn costs per meter-cubed

I repeated the same method I used with the 5-D polynomial and determined that the following fifth-order polynomial best approximated the data:

$$C_5(d,r) = \beta_1 d^5 + \beta_2 d^4 r + \beta_3 d^4 + \beta_4 d^3 r^2 + \beta_5 d^3 r + \beta_6 d^3 + \beta_7 d^2 r^3 + \beta_8 d^2 r^2 + \beta_9 d^2 r + \beta_{10} d^2 + \beta_{11} d r^4 + \beta_{12} d r^3 + \beta_{13} d r^2 + \beta_{14} d r + \beta_{15} d + \beta_{16} r^5 + \beta_{17} r^4 + \beta_{18} r^3 + \beta_{19} r^2 + \beta_{20} r + \beta_{21}$$

N	Coefficient β_n
1	7.12033097648382E-03
2	1.99441192077618E-04
3	-1.83101314497891E+00
4	-3.91919219460648E-05
5	1.41487331357488E-01
6	-1.62444332147031E+01
7	8.28224229423360E-07
8	-1.71096427025266E-03
9	-2.96576494728825E+00
10	2.17445571750448E+03

N	Coefficient β_n
11	5.28560972447234E-09
12	-7.71000033652331E-05
13	2.44901711876568E-01
14	-2.14442923320968E+02
15	4.90406404525024E+04
16	1.68493005344143E-10
17	-1.80361479242544E-06
18	7.95473848140804E-03
19	-1.70936117838247E+01
20	1.70178164664643E+04
21	-6.25905106725071E+06

Table 6 Coefficients of *$C_5(d,r)$*

This polynomial had the least error of the attempted approximations, with and RMSE of 1555.49095458590.

Because of the limitations of using a high-order polynomial (discussed in *Limitations and Recommendations*), I also approximated the data using a third-order polynomial:

$$C_3(d,r) = \gamma_1 d^3 + \gamma_2 d^2 r + \gamma_3 d^2 + \gamma_4 d r^2 + \gamma_5 d r + \gamma_6 d + \gamma_7 r^3 + \gamma_8 r^2 + \gamma_9 r + \gamma_{10}$$

N	Coefficient γ_n
1	1.60785249816796E-01
2	-9.67853886666766E-03
3	6.18204392443374E+00
4	-2.98502603295663E-04
5	1.92029273329078E+00
6	-2.39236650599339E+03
7	8.87468360175106E-06
8	-5.28650499314268E-02
9	9.16445204287303E+01
10	-3.81802502457326E+04

Table 7 Coefficients of *$C_3(d,r)$*

Test Cases

Again, I tested these models by 'predicting' the outturn costs of the tunnels used in the dataset, as well outturn costs of other tunnels in the database not used in the dataset, chosen at random.

Data Fit

5th Order Polynomial

The results of the outturn cost cost-prediction test of tunnels used in the dataset:

Depth (m)	Substrate Density (kg/m^3)	Tunnel Outturn cost per m^3 (USD)	$C_5(d,r)$ Prediction of Outturn cost per m^3 (USD)	Absolute Error
39.5	2067.5	12195.38426	12095.3225	100.0617559
70	2450	4492.286503	4492.646876	0.360373079
27.4	2050	4768.737163	5103.861868	335.1247048
20	2365	10811.6381	6669.796529	4141.841571
35	2265	3096.679501	2376.258721	720.4207805
17.5	2615	4727.086566	4674.449884	52.63668217
30	2450	2855.530544	2237.437723	618.0928213
21	1520	8741.892776	8491.746106	250.1466699
30	1806.7	6578.300307	3229.574007	3348.7263
30	1806.7	5.018031401	3229.574007	3224.555976
30	1806.7	3513.48221	3229.574007	283.9082035
20	2365	2408.386663	6669.796529	4261.409867
53	2625	1908.534399	1887.655243	20.87915598
17.5	1200	1972.330202	1928.096886	44.23331626
10.75	1520	12990.19608	13106.95502	116.7589385
65	1907.5	11595.3001	11590.28508	5.01502188
30	1.675	1108.377498	2066.446421	958.0689228
45	1675	811.7880213	845.0773205	33.28929919
30	1550	685.9770922	110.9283124	575.0487798
60	2300	2368.069087	2381.769954	13.70086755
35	2365	837.9943068	1858.701868	1020.707561
15	1835	734.6554561	591.8447963	142.8106598
2	1200	7896.685097	7886.041001	10.64409646
24	2615	1497.169423	1630.002809	132.8333861
22.6	1333.3	766.3840414	984.0399654	217.6559239
			Average error	825.1572653

Table 8 Comparison between actual and modeled outturn cost costs per m^3 of tunnels in the dataset using the 3-dimensional polynomial

The predictions were not as accurate as those of $O(f,l,d,r)$, having an average absolute error of $825.16. However, this error may be acceptable given the inherent uncertainty of outturn cost estimation during early stages of tunnel delivery.

A 3-D plot of the polynomial surface is shown below, with a scatter plot of the dataset (green points) included to visualize the model's data fit:

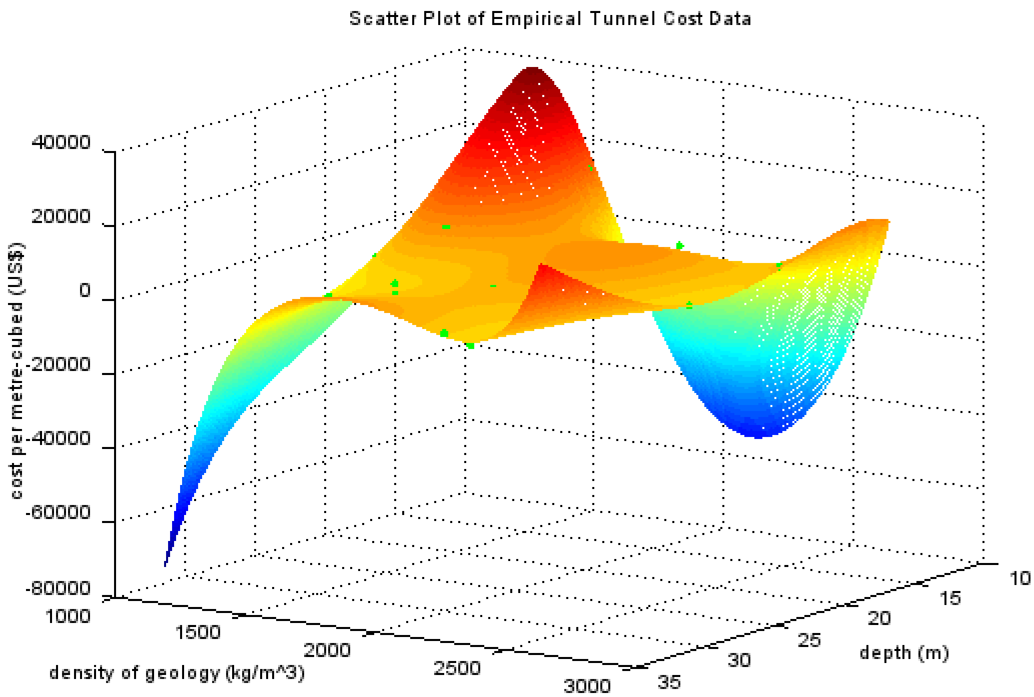

Figure 1 close-up of the surface generated by the 5th order polynomial $C_5(d,r)$ demonstrating the data fit

3rd Order Polynomial

The results of the outturn cost-prediction test of tunnels used in the dataset:

Depth (m)	Substrate Density (kg/m^3)	Tunnel Outturn cost per m^3 (USD)	$C_3(d,r)$ Prediction of Outturn cost per m^3 (USD)	Absolute Error
39.5	2067.5	12195.38426	4008.95039	8186.433865
70	2450	4492.286503	5230.282797	737.9962933
27.4	2050	4768.737163	4974.910111	206.1729483
20	2365	10811.6381	4460.90684	6350.731261
35	2265	3096.679501	3820.69396	724.0144594
17.5	2615	4727.086566	3957.24238	769.8441855
30	2450	2855.530544	3720.096479	864.5659349
21	1520	8741.892776	4446.968424	4294.924352
30	1806.7	6578.300307	4418.314988	2159.985319
30	1806.7	5.018031401	4418.314988	4413.296957
30	1806.7	3513.48221	4418.314988	904.8327776
20	2365	2408.386663	4460.90684	2052.520177
53	2625	1908.534399	-74.4234569	1982.957856
17.5	1200	1972.330202	1138.519939	833.8102634
10.75	1520	12990.19608	7606.059899	5384.136179
65	1907.5	11595.3001	10138.75923	1456.540871
30	1.675	1108.377498	3624.465623	2516.088125
45	1675	811.7880213	2451.184344	1639.396323
30	1550	685.9770922	2320.2077	1634.230608

60	2300	2368.069087	4484.658806	2116.58972
35	2365	837.9943068	3477.443293	2639.448986
15	1835	734.6554561	6645.597481	5910.942025
2	1200	7896.685097	9946.795153	2050.110056
24	2615	1497.169423	3980.442229	2483.272805
22.6	1333.3	766.3840414	1292.270004	525.8859623
			Average Error	2513.549132

Table 9 Comparison between actual and modeled outturn costs per m^3 of tunnels in the dataset using the simplified 3-dimensional polynomial

Again, the predictions were not as accurate as those of *O(f,l,d,r)*, having an average absolute error of $2513.55. However, this error may be acceptable given the inherent uncertainty of outturn cost estimation during early stages of tunnel delivery.

A 3-D plot of the polynomial surface is shown below, with a scatter plot of the dataset (green points) included to visualize the model's data fit:

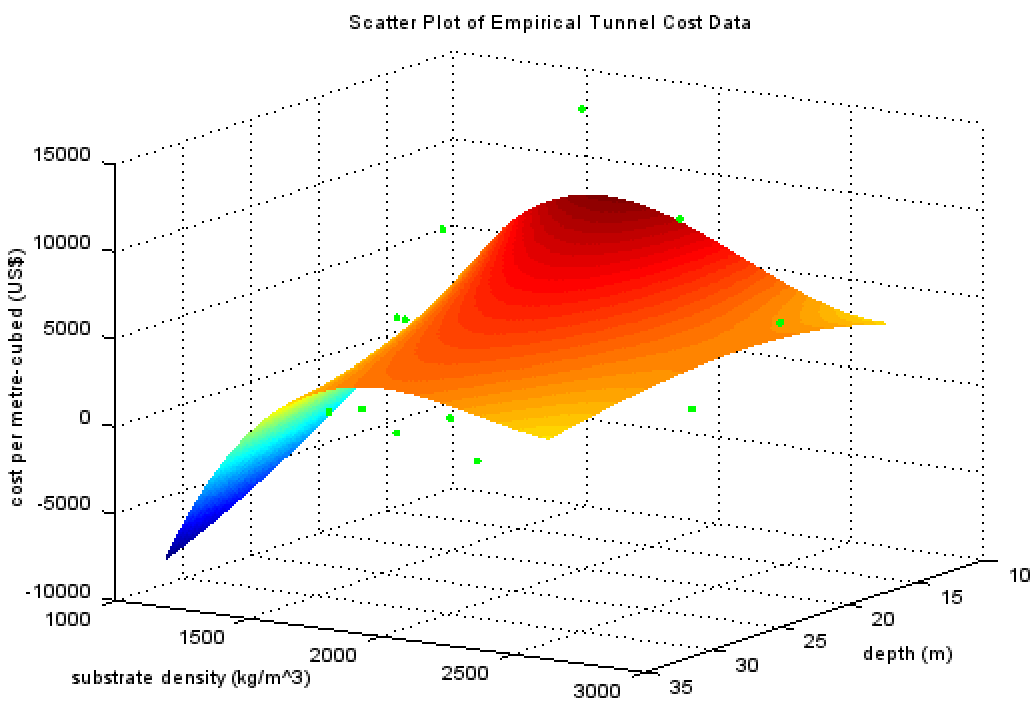

Figure 2 A close-up of the surface generated by the $C_3(d,r)$ demonstrating data fit

Prediction Test

The results of the outturn cost-prediction test of tunnels not used in the dataset:

5th-Order Polynomial

Tunnel	Depth (m)	Substrate Density (kg/m^3)	Tunnel Outturn cost per m^3 (USD)	*$C_5(d,r)$* Prediction of Outturn cost per m^3 (USD)	Absolute Error
UK Tunnel 9	21	1600	7486.528961	7783.613101	297.0841396
University Link (ULINK) Project	22.1	1573.3	8741.892776	6259.180347	2482.712428
Dulles Airport Train System	15	1935	29865.44419	-2355.764242	32221.20843
UK Tunnel 1	21	1760	546.5202553	6647.541746	6101.021491
Sants-Sagrera HSL	34	1985	481.8556235	6627.041331	6145.185708
Channel Tunnel	45	1825	811.7880213	22687.41876	21875.63074
Bjørvika Immersed Tunnel	9	1200	3520.946099	36478.79182	32957.84572
				Average Error	14582.95552

Table 10 Comparison between actual and modeled outturn costs per m^3 of tunnels chosen at random using the 3-dimensional polynomial

Despite the tolerable predictions for UK Tunnel 9 and ULINK, the predictions are inaccurate with an average absolute error of $14582.96. However, taking the worst case of error from this set, the Bjørvika Immersed Tunnel, this model's prediction has an error of $32957.85 in outturn cost per meter cubed. Multiplying this error by the volume of the tunnel (243508 m^3), the error in the raw outturn cost is $8025499096, which is three orders of magnitude less than the error of *O(f,l,d,r).* This shows that $C_5(d,r)$ may be a better predictor of outturn cost than *O(f,l,d,r),* if still inaccurate. Again, note the negative outturn cost prediction.

3rd-Order Polynomial

Tunnel	Depth (m)	Substrate Density (kg/m^3)	Tunnel Outturn cost per m^3 (USD)	*$C_5(d,r)$* Prediction of Outturn cost per m^3 (USD)	Absolute Error
UK Tunnel 9	21	1600	7486.528961	5087.934453	2398.594508
University Link (ULINK) Project	22.1	1573.3	8741.892776	4595.469485	4146.423291
Dulles Airport Train System	15	1935	29865.44419	6316.908417	23548.53577
UK Tunnel 1	21	1760	546.5202553	5762.180805	5215.660549
Sants-Sagrera HSL	34	1985	481.8556235	4373.301399	3891.445776
Channel Tunnel	45	1825	811.7880213	3651.618313	2839.830292
Bjørvika Immersed Tunnel	9	1200	3520.946099	6019.428885	2498.482787
				Average Error	6362.710424

Table 11 Comparison between actual and modeled outturn costs per m^3 of tunnels chosen at random using the simplified 3-dimensional polynomial

$C_3(d,r)$seems to be more accurate a predictor than $C_5(d,r)$, with an average absolute error of $6362.71. It also appears to be more stable than $C_5(d,r)$, which makes it appear to be more reliable despite having a higher RMSE.

Data Imputation

I used each model to make a prediction of the outturn costs and outturn costs per meter-cubed for three tunnels in the database that have input data but not recorded outturn cost data.

Tunnel	Face area (m^2)	Total tunnel length (m)	Depth (m)	Substrate Density (kg/m^3)	*O(f,l,d,r)* Prediction of Outturn cost (USD)	*C_5(d,r)* Prediction of Outturn cost per m^3 (USD)	*C_5(d,r)* Prediction of Outturn cost per m^3 (USD)
Confidential at present	108.61868	5899	192.5	2800	1.73248519748721E+13	6.02807741150758E+08	494582.8852
Vector Cable Tunnel	9.6211275	9000	80	2700	-2.85905407729007E+09	2.84114093802599E+04	2503.307321
OARS (Olentangy-Scioto Intercepting Sewer Augmentation Relief Sewer)	35.361845	7242	36.6	2650	-4.09410200772199E+10	3.36545029155118E+04	2360.81872

Table 12 Outturn cost Imputation of three tunnels from the database

It is possible to use outturn costs and three inputs to impute the fourth. However, these models may be too inaccurate to perform that procedure.

Conclusion

I obtained the following model for raw tunnel outturn cost in terms of face area *f,* length *l,* depth *d,* and density *r*:

$$O(f,l,d,r) = \alpha_1 f^3 + \alpha_2 f^2 l + \alpha_3 f^2 d + \alpha_4 f^2 r + \alpha_5 f^2 + \alpha_6 f l^2 + \alpha_7 fld + \alpha_8 flr + \alpha_9 fl + \alpha_{10} fd^2 + \alpha_{11} fdr + \alpha_{12} fd + \alpha_{13} fr^2 + \alpha_{14} fr + \alpha_{15} f + \alpha_{16} l^3 d + \alpha_{17} l^2 d + \alpha_{18} l^2 r + \alpha_{19} l^2 + \alpha_{20} ld^2 + \alpha_{21} ldr + \alpha_{22} ld + \alpha_{23} lr^2 + \alpha_{24} lr + \alpha_{25} l + \alpha_{26} d^3 + \alpha_{27} d^2 r + \alpha_{28} d^2 + \alpha_{29} dr^2 + \alpha_{30} dr + \alpha_{31} d + \alpha_{32} r^3 + \alpha_{33} r^2 + \alpha_{34} r - \alpha_{35}$$

I also found a slightly better model for outturn cost per meter-cubed in terms of depth and density:

$$C_5(d,r) = \beta_1 d^5 + \beta_2 d^4 r + \beta_3 d^4 + \beta_4 d^3 r^2 + \beta_5 d^3 r + \beta_6 d^3 + \beta_7 d^2 r^3 + \beta_8 d^2 r^2 + \beta_9 d^2 r + \beta_{10} d^2 + \beta_{11} dr^4 + \beta_{12} dr^3 + \beta_{13} dr^2 + \beta_{14} dr + \beta_{15} d + \beta_{16} r^5 + \beta_{17} r^4 + \beta_{18} r^3 + \beta_{19} r^2 + \beta_{20} r + \beta_{21}$$

and simplified it into a slightly better model still:

$$C_3(d,r) = \gamma_1 d^3 + \gamma_2 d^2 r + \gamma_3 d^2 + \gamma_4 dr^2 + \gamma_5 dr + \gamma_6 d + \gamma_7 r^3 + \gamma_8 r^2 + \gamma_9 r + \gamma_{10}$$

Coefficients α_n, β_n, and γ_n can be found in the *Results* section.

Limitations and Recommendations

With my work came several limitations, both inherent and imposed.

Data

The data I used came from a database developed at AECOM. The information on each tunnel is basic and was gathered mostly from public reports on the Internet. The data is somewhat unreliable and almost always rounded to the two or three significant figures, making them imprecise. Also, different and sources sometimes gave different data on the tunnels, and judgment was used to decide which to 'trust'. This added to the inaccuracy of the data.

Much of the data on the tunnels was qualitative, such as the geology, the location of the tunnel, and so on. I only used data that I could quantitatively characterize, which excluded tunnel location, end use and setting. Unfortunately, this meant I could not represent each tunnel comprehensively in the dataset. For example, I did not include tunnel setting in my analysis. However, it is well known that tunnels constructed in urban areas have a higher outturn cost than those in rural areas because of the enabling works necessary to ensure little disruption of civilian life during and after construction. Exclusion of this factor from my model weakens its predictive power.

I was able to quantify only the geology. I considered several methods of representing the 'essence' of the geology as a number, including penetration resistances (result of the cone penetration test, a standard geotechnical investigation technique) [4] and rippability indices (measure of excavatability) [8]. However, as far as I could determine, rippability properties of rock are not of much concern to conventional tunnel boring technology, but rather bulldozing [6]. Also, penetration tests return widely different results for general categories of ground (mudstone, sandstone, and so on) [5]. This is because several factors affect the penetrability of the ground, including saturation and weathering. This means penetration measures are specific to each site [7]. Because the database did not feature such detail, I decided to use a simple measure of the average density of geology. The average geological density does not capture the factors mentioned about (weathering, etc.) that affect excavatability, but is more easily available and less variable among sites.

Also, the database is incomplete; there are large gaps in data. Of the 162 tunnels in database, only 25 of them had sufficient data for my methods. This reduction in our sample size lowered our models potential to capture empirical effects, and thus lowered its prediction power.

Recommendations for Data Limitations

I recommend that members of the tunnel industry worldwide work toward centralizing accurate tunnel data. I also recommend that this data be comprehensive and specific in a way such that is can be quantitatively characterized. For example, the cost of labor, the quantitative results of geotechnical investigation and a measure of the extent of enabling works need to facilitate tunnel construction. All this information will make numerical analysis on empirical data more accurate and more fruitful. This is important because the ability to accurately estimate tunnel outturn cost will benefit both clients and contractors worldwide.

Methods

I treated empirical tunnel data abstractly to develop a polynomial approximation that represents tunnel outturn costs in terms of basic tunnel characteristics. This has three effects which limit the usefulness of the model:

1. The model polynomial created through multi-dimensional approximation is itself abstract. Each input is treated independently of the others. It does not express any obvious real world relations between the factors investigated and offers little insight into these relations. In fact, it is possible that the inputs used are not independent in such a way that a polynomial created by my method can be accurate. For example, the depth of a tunnel may not be independent of, say, its length, in the sense that a deeper tunnel is more/less expensive regardless of other factors.

2. The data used did not adequately capture the effect of indirect factors that vary by country, for example, the cost of labor involved in tunnel construction. Thus, it is possible for two tunnels to have similar inputs (as defined in the dataset) but very different outturn costs because of these factors. This limits the power of the polynomial approximation because of the smooth and continuous nature of polynomials; resulting polynomials might 'bounce around', which make them unreliable. This occurs most frequently with high-order polynomials.

3. The outputs of the model must be interpreted abstractly. Certain combinations of inputs may yield negative outturn costs, as seen in the *Results* section. Negative costs in this context are nonsensical and do not help the estimation process.

It is apparent that polynomial approximation is not the best method to use to create a model in this situation. I chose the method used because of limitations in time, limitations in access to information on tunnel construction, and limited understanding in general of the exact effect of all the factors on tunnel cost. (I was also most comfortable with polynomial approximation as a method for this type of analysis.)

Recommendations for Methodological Limitations

As mentioned before, the main problem polynomial approximation in this context is the empirical data, as represented in my dataset, is erratic and 'noisy'. Thus, it is very difficult to obtain a smooth and robust polynomial which accurately approximates the data.

I recommend methods that reduce the noisiness of the data so that more *reliable* approximations can be done. I recommend factor analysis can be done to determine whether there are any unobserved factors that relate the characteristics of tunnels. If these unobserved factors can be determined, then my method can be applied to those factors for more accuracy and possible real-world insight (addressing problem 1 above). I also recommend statistical methods be used to remove outlier data. Removing extreme data points could allow this method to produced more stable polynomials, which will likely result in a more reliable prediction.

However, this may still subject the resulting model to the limitations discussed above. To that end, I recommend spline interpolation. This method will approximate the data with special piecewise polynomials known as splines. Using splines instead of a single polynomial will reduce approximation error and will more likely produce a more accurate and more reliable predictive model.

In the end, I believe that best results will be obtained through further research into the effects of various factors on tunnel construction to build a bottom-up model. This will allow insight into the relationship between factors that affect tunnel outturn costs.

References

[1] A. Alden, "Density of Common Rock Types," vol. 2012, http://geology.about.com/cs/rock_types/a/aarockspecgrav.htm.

[2] ChemPRIME, "Density of Rocks and Soils," *ChemPRIME,* vol. 2012, http://wiki.chemprime.chemeddl.org/index.php/Density_of_Rocks_and_Soils .

[3] J. D'Errico, "File Exchange - polyfitn," vol. 2012, .

[4] Geotechdata.info, "Cone Penetration test," vol. 2012, 29 November 2010.

[5] P. K. Robertson, "Soil Classification using the Cone Penetration Test," October 13, 1989.

[6] I. Assakkaff, "Construction Planning Equipment and Methods: Dozers," February 23, 2003.

[7] R. K. Goel and B. Singh, "Method of excavation," in *Engineering Rock Mass Classification: Tunnelling, Foundations, and Landslides*Anonymous Oxford, UK: Elsevier, 11, pp. 281-283.

[8] P. Quigley and S. McSwiney, "MEASUREMENT AND ANALYSIS OF ROCK MASS FRACTURES AND THEIR APPLICATIONS IN CIVIL ENGINEERING," 29 November 2010, 2010.

[9] J. Koloski, S. Schwarz and D. Tubbs, "Geotechnical Properties of Geologic Materials," vol. 2012, Novermber 26, 2010, 2010.