Predicting costs :: Part 2

Planning and Regional Development

1/21/2020

Large projects - defined as the biggest 20 percent of 239 projects evaluated in Part 1 — and ones subjected to closed bidding processes explain a significant degree of the inaccuracy observed in agency cost estimation. Part 2 searches for predictive factors from within those two project subpopulations. It also considers the relationship between accuracy and the number of bidders, which was generally omitted from Part 1 as internal estimators do not know how many bidders will respond as they develop estimates.

The agency's internal cost estimates predict 95 percent of variation in cost, using the second-lowest bid[1] as a predicting target. On an absolute basis[2], however, the gap between internal estimate and second-lowest bid averages $2.7 million, or 18 percent of the average project size. Reducing this gap would provide for stronger confidence in long-range capital capacity estimates and could reduce the need for project-level change orders.

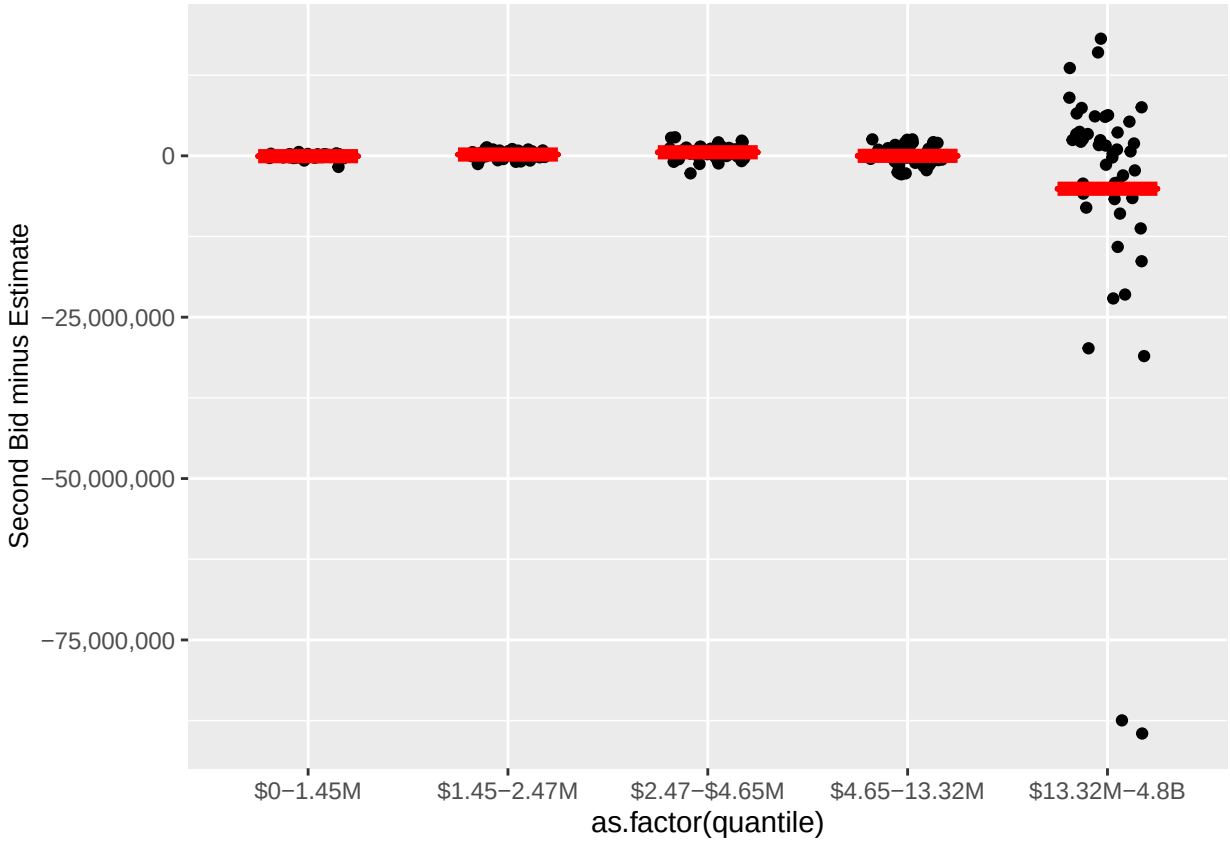
2. Motivation

```
bids$Format = relevel(bids$Format, ref = "Public")
```

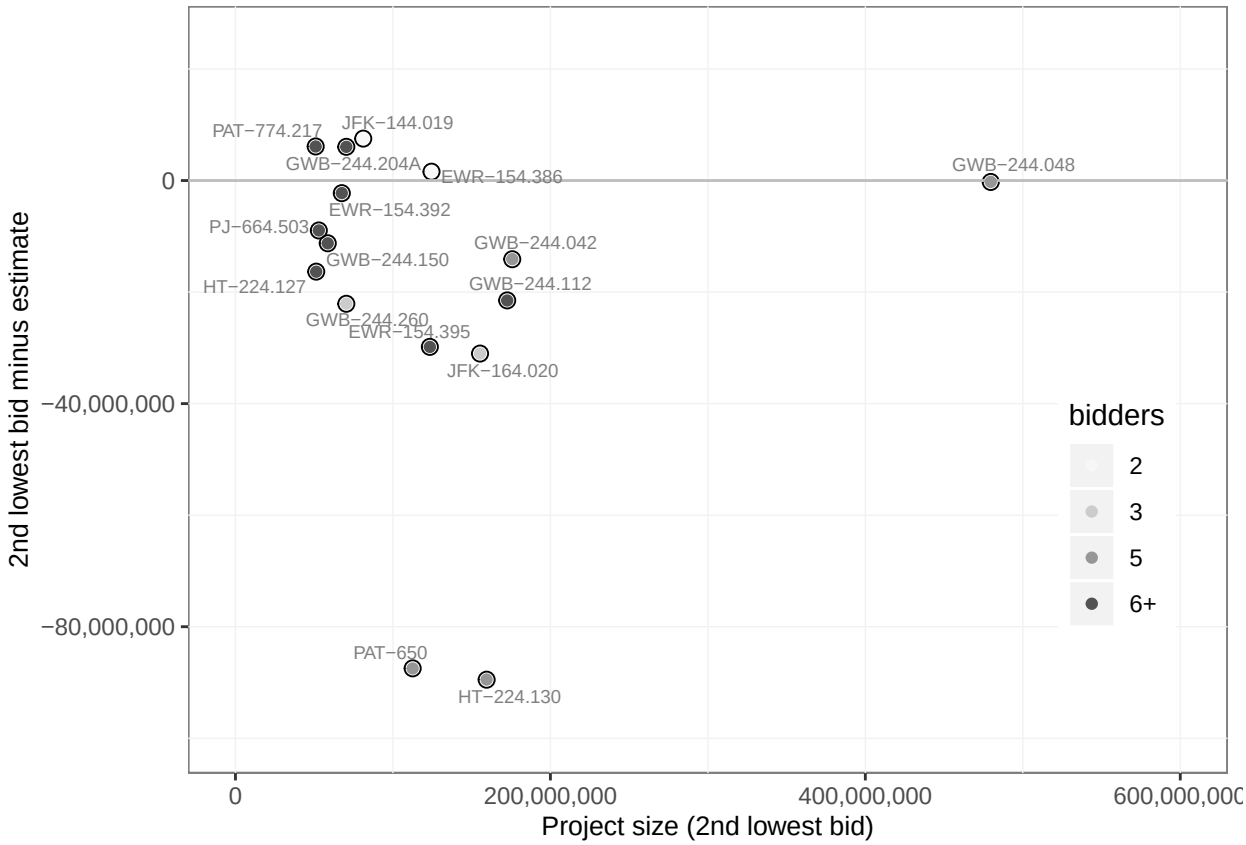
[1]The agency's internal regulations require, in all but a handful of cases, the acceptance of the lowest bid. The Engineering Department views the second-lowest bid as a better predictive target.

[2]Mean absolute error, MAE.

Project size (dollars bid)




```
bids.big = subset(bids, bids$Second.Bid >= 5e+07)
```



Internal bids for the largest projects are too high. There is no observed systemic relationship here between that inaccuracy and the number of bidders (bids) per project.

Bidding process

Big projects don't have any SBE or M/WBE presence, but there are quite a few non-public processes. They don't appear to have much influence on accuracy at that level:

```
summary(lm(bids.big$accuracy ~ bids.big$Format))
```

```
##
## Call:
## lm(formula = bids.big$accuracy ~ bids.big$Format)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38628 -0.05461 -0.00176  0.07616  0.24494
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.84897    0.05080  16.712 1.21e-10 ***
## bids.big$FormatOther  0.10000    0.08295   1.205    0.248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1606 on 14 degrees of freedom
```

```
## Multiple R-squared:  0.09403,    Adjusted R-squared:  0.02932
## F-statistic: 1.453 on 1 and 14 DF,  p-value: 0.248
```

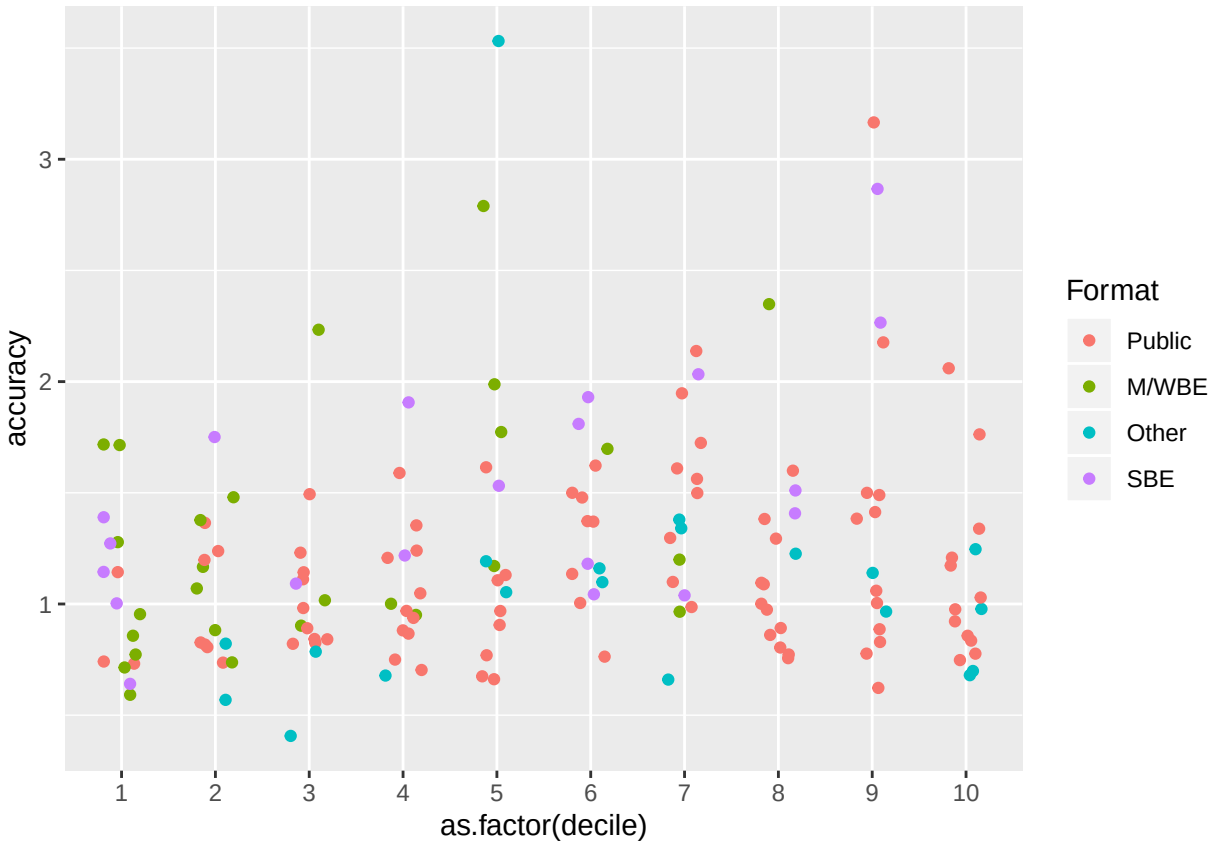
```
summary(lm(bids.big$bal ~ bids.big$Format))
```

```
##
## Call:
## lm(formula = bids.big$bal ~ bids.big$Format)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -72385618  -1744306   8492888  16413632  28326996
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -22283246    9615067  -2.318   0.0361 *
## bids.big$FormatOther   7205865   15701338   0.459   0.6533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30410000 on 14 degrees of freedom
## Multiple R-squared:  0.01482,    Adjusted R-squared:  -0.05555
## F-statistic: 0.2106 on 1 and 14 DF,  p-value: 0.6533
```

The statistical significance of bidding process identified earlier was limited to smaller projects. Was it the difference between public processes and SBE-slash-M/WBE? What about the "other"[3] category?

```
##
##           1  2  3  4  5
##   Public 20 27 33 34 27
##   M/WBE  17  7  3  0  0
##   Other   4  7  6 14 20
##   SBE     7  7  6  0  0
```

[3]This included security projects and other non-descript processes




```
##  Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 1 9 3 8 8 4 4 5 1 5 ...
```

```
##    Group.1       x
## 1        1  686000
## 2        2 1047850
## 3        3 1444444
## 4        4 1647000
## 5        5 1992121
## 6        6 2440000
## 7        7 2765000
## 8        8 3769000
## 9        9 4483190
## 10      10 5887300
```

There isn't much of an issue for the very smallest projects. Estimates are pretty much on the money, irrespective of the type of bidding process involved. And - they're very small projects, even more justification to focus on remaining projects, which fall between $1 million and $50 million in size.

```
summary(lm(bids.small$accuracy ~ bids.small$Format))
```

```
##
## Call:
## lm(formula = bids.small$accuracy ~ bids.small$Format)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7165 -0.3434 -0.1090  0.2171  2.4083
##
## Coefficients:
```

```
##                         Estimate Std. Error t value             Pr(>|t|)
## (Intercept)              1.16856    0.05437  21.493 < 0.0000000000000002 ***
## bids.small$FormatM/WBE   0.34602    0.14311   2.418              0.01705 *
## bids.small$FormatOther -0.04522    0.12879  -0.351              0.72611
## bids.small$FormatSBE     0.46249    0.14311   3.232              0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4953 on 125 degrees of freedom
## Multiple R-squared:  0.1116, Adjusted R-squared:  0.09027
## F-statistic: 5.234 on 3 and 125 DF,  p-value: 0.001951
```

The implications of bidding process emerges from the difference between average accuracy for projects bid publicly and those tagged SBE and M/WBE. The relationship loses statistical significance when turning to public-versus-M/WBE projects, but not by much, and the coefficent is the same sign, and we lump SBE and M/WBE together for a collective look:

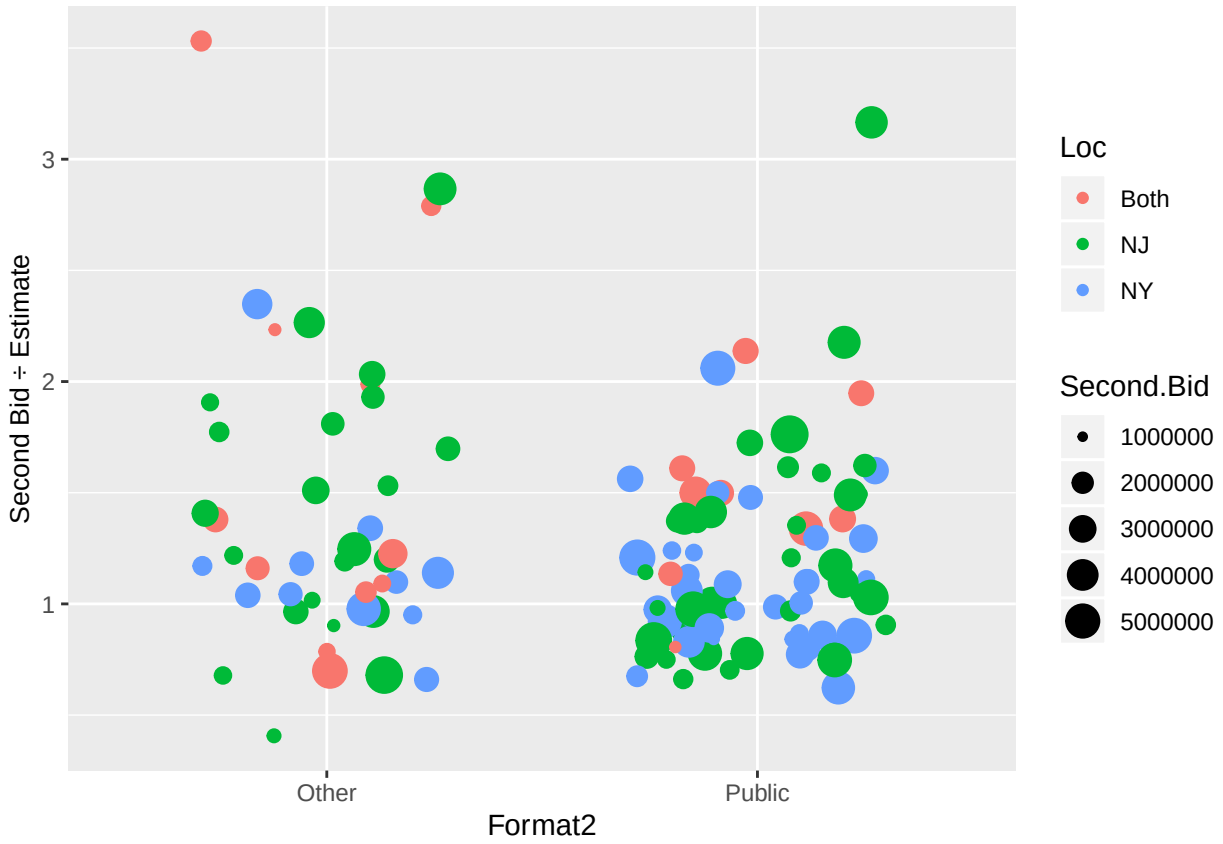
```
summary(lm(bids.small$accuracy ~ bids.small$Format2))
```

```
##
## Call:
## lm(formula = bids.small$accuracy ~ bids.small$Format2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9901 -0.3394 -0.1643  0.2448  2.1347
##
## Coefficients:
##                           Estimate Std. Error t value            Pr(>|t|)
## (Intercept)                1.39694    0.07513  18.593 <0.0000000000000002 ***
## bids.small$Format2Public -0.22838    0.09367  -2.438              0.0161 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5096 on 127 degrees of freedom
## Multiple R-squared:  0.04472,    Adjusted R-squared:  0.03719
## F-statistic: 5.945 on 1 and 127 DF,  p-value: 0.01614
```

So what could be explaining the influence bidding process has on accuracy? First explain bidding process. Does it vary significantly by project size? Project typeology?[4] By line department or location?

```
ggplot(bids.small, aes(x = Format2, y = accuracy, color = Loc, size = Second.Bid)) +
    geom_hline(colour = "dark gray", yintercept = 1) + geom_jitter(width = 0.3) +
    ylab("Second Bid ÷ Estimate") + theme(axis.title.y = element_text(size = 10))
```

[4](If so, it could help explain the accuracy's variation (in Part 1) by typeology.)




```
summary(glm(bids.small$Format2 ~ bids.small$decile + bids.small$Typeology + bids.small$Loc +
    bids.small$LD, family = binomial))
```

```
##
## Call:
## glm(formula = bids.small$Format2 ~ bids.small$decile + bids.small$Typeology +
##     bids.small$Loc + bids.small$LD, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2667  -0.9957   0.5202   0.8139   1.5653
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.1043     0.9108   0.114   0.9089
## bids.small$decile.L           0.5786     0.9382   0.617   0.5374
## bids.small$decile.Q           0.9310     0.9477   0.982   0.3259
## bids.small$decile.C           0.1700     0.8885   0.191   0.8483
## bids.small$decile^4          -0.6410     0.7514  -0.853   0.3936
## bids.small$decile^5           0.6570     0.6889   0.954   0.3402
## bids.small$decile^6           0.5111     0.6495   0.787   0.4314
## bids.small$decile^7          -0.4253     0.6079  -0.700   0.4842
## bids.small$decile^8           0.3672     0.5907   0.622   0.5342
## bids.small$TypeologyInfra     1.0893     0.5449   1.999   0.0456 *
## bids.small$TypeologyPaving   -1.1423     0.6017  -1.898   0.0577 .
## bids.small$LocNJ              0.2287     0.7107   0.322   0.7476
## bids.small$LocNY              1.0956     0.7858   1.394   0.1633
```

```
## bids.small$LDPATH           -0.3153     0.7266  -0.434   0.6643
## bids.small$LDPorts          -0.4302     0.6871  -0.626   0.5313
## bids.small$LDTB&T           -0.1620     0.6423  -0.252   0.8009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 168.07  on 128  degrees of freedom
## Residual deviance: 139.46  on 113  degrees of freedom
## AIC: 171.46
##
## Number of Fisher Scoring iterations: 4
```

```
ggplot(bids.small, aes(x = as.factor(decile), y = accuracy, colour = Bids)) + scale_colour_gradient(low
    high = "black") + geom_jitter()
```