QTP
Quality Transport Planning

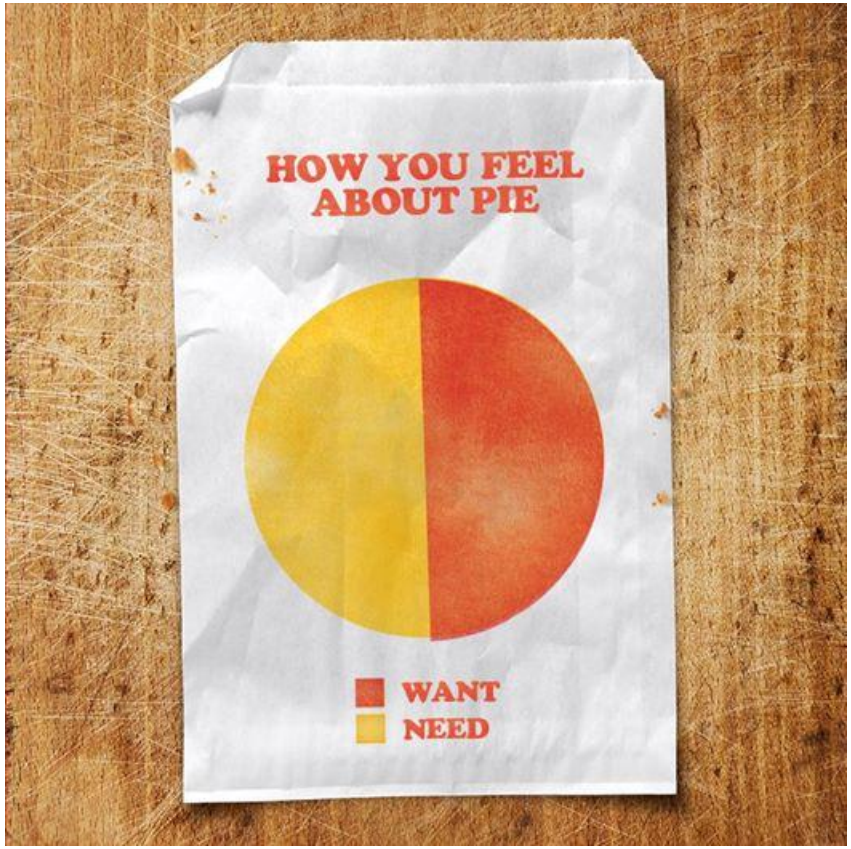
Model Measures: Are *they* fit for purpose?

Tim Wright

NZMUGS Conference, 2014

Introduction

- ‘Fit for Purpose’
- Draft MUGS Data Comparison Guidelines (GEH)
- Expanded review of:
  - GEH
  - R^2
  - RMSE
- I am no statistician (but I love statistics !)
- ‘Keeping it real’



“For every grain of sand on Earth, there are 10,000 stars in the observable universe”

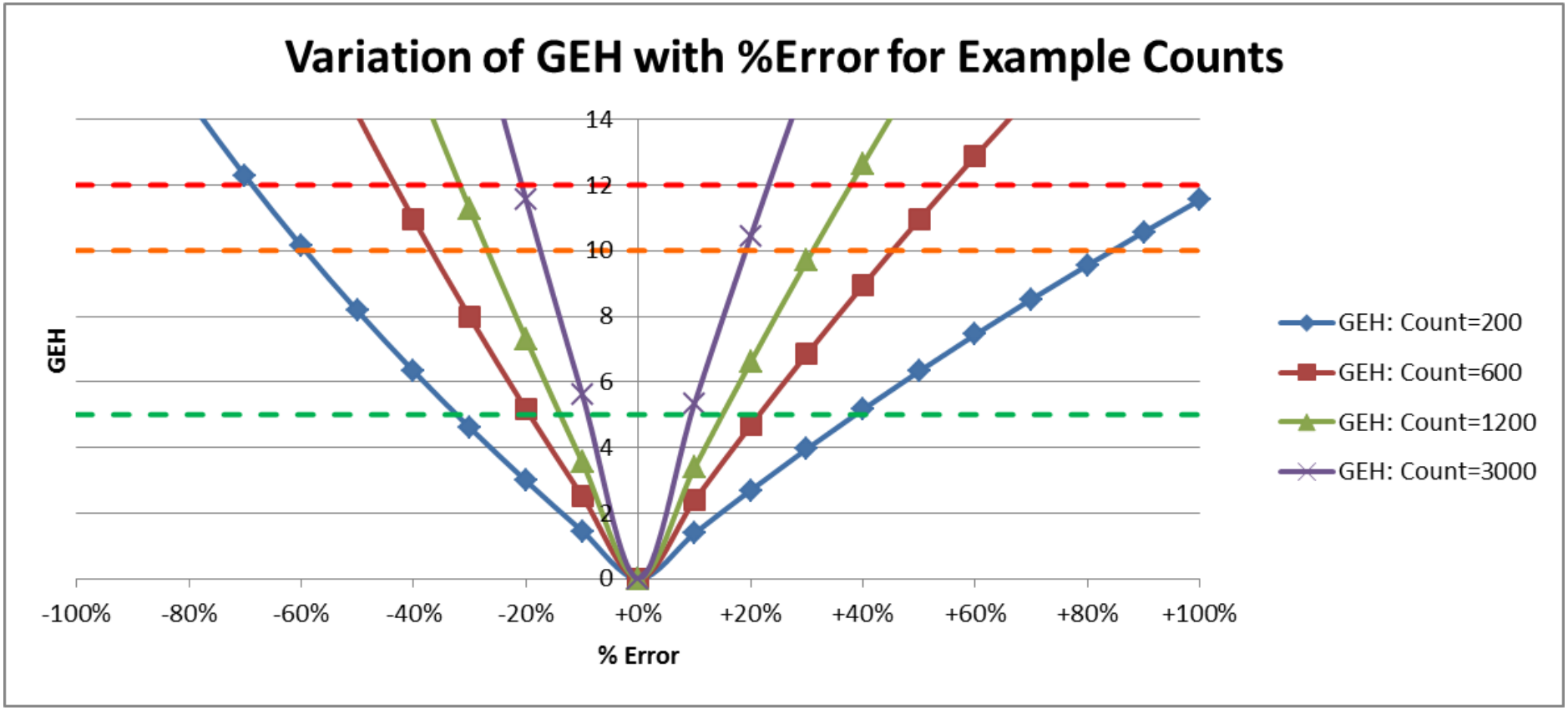
“1 Million seconds is around 12 Days. 1 Billion seconds is around 32 years”

GEH Definition

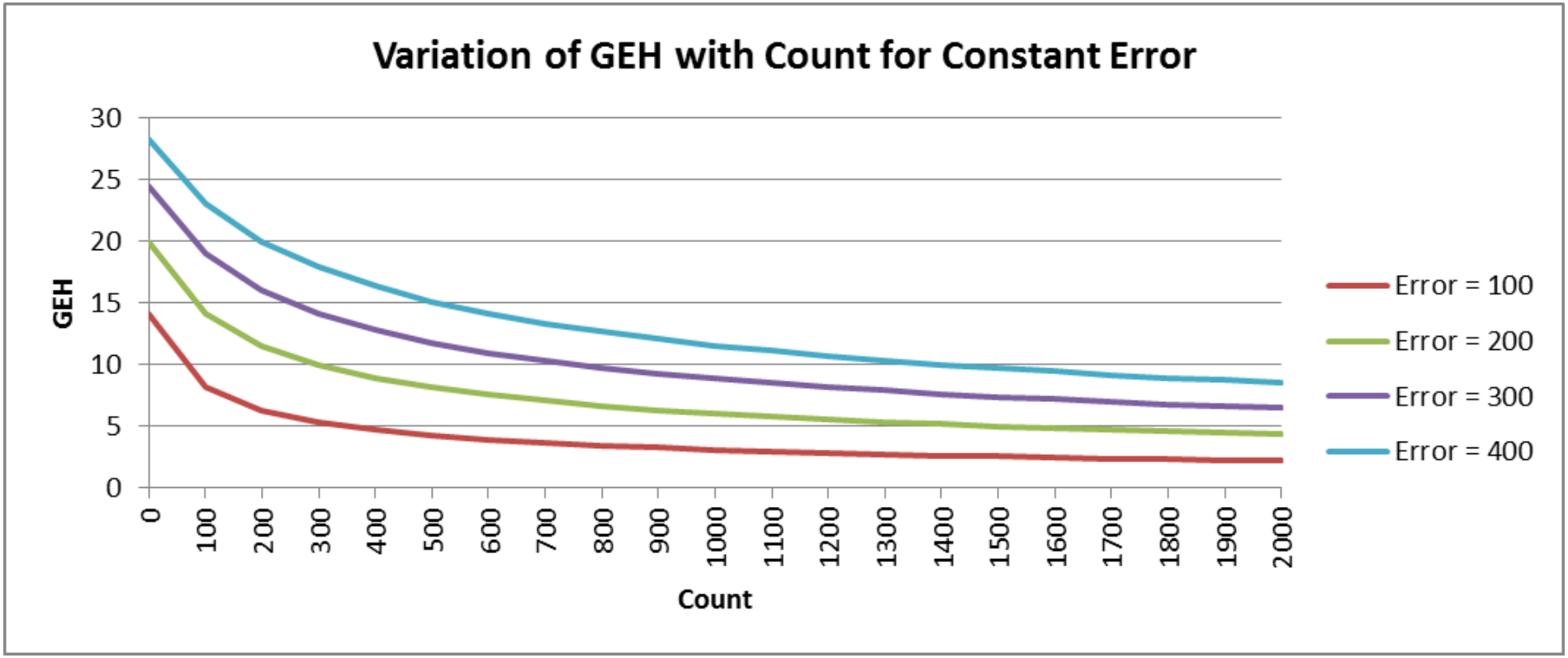
- $\sqrt{\frac{(M-O)^2}{(M+O)/2}}$
- $\frac{\sqrt{(M-O)^2}}{\sqrt{(M+O)/2}}$
- $\frac{(M-O)}{\sqrt{(M+O)/2}}$
- $\frac{Difference}{\sqrt{Average}}$

GEH Trend

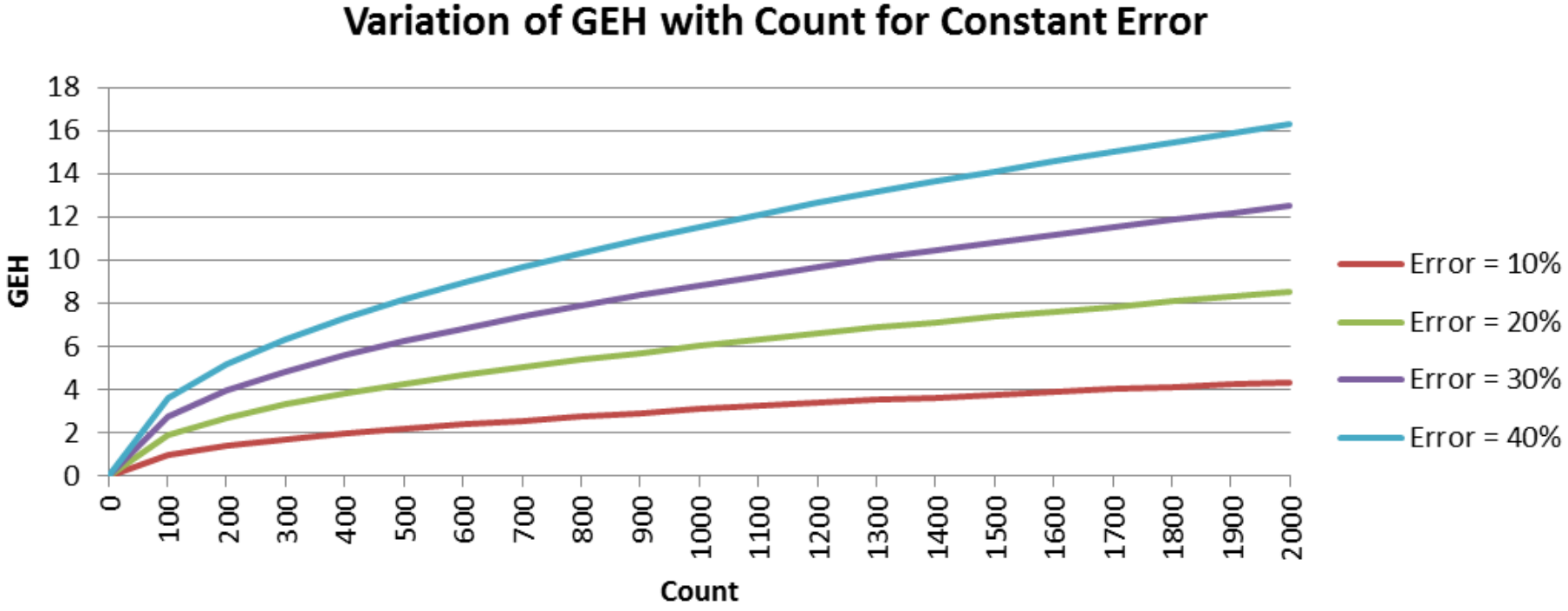
- Single measure of overall fit



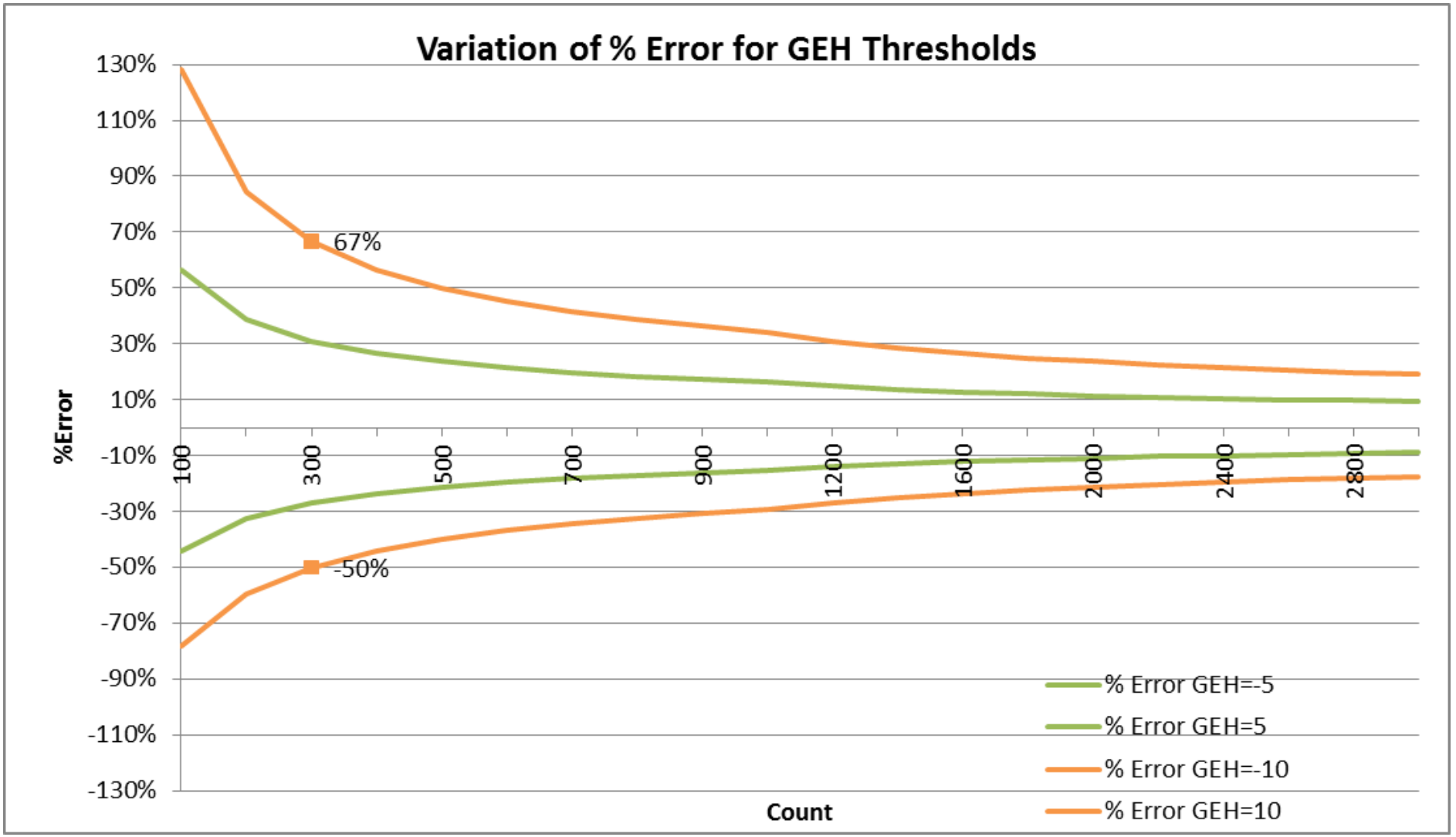
GEH Trend


Variation of GEH with Count for Constant Error
GEH
30
25
20
15
10
5
0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
Count
Error = 100
Error = 200
Error = 300
Error = 400

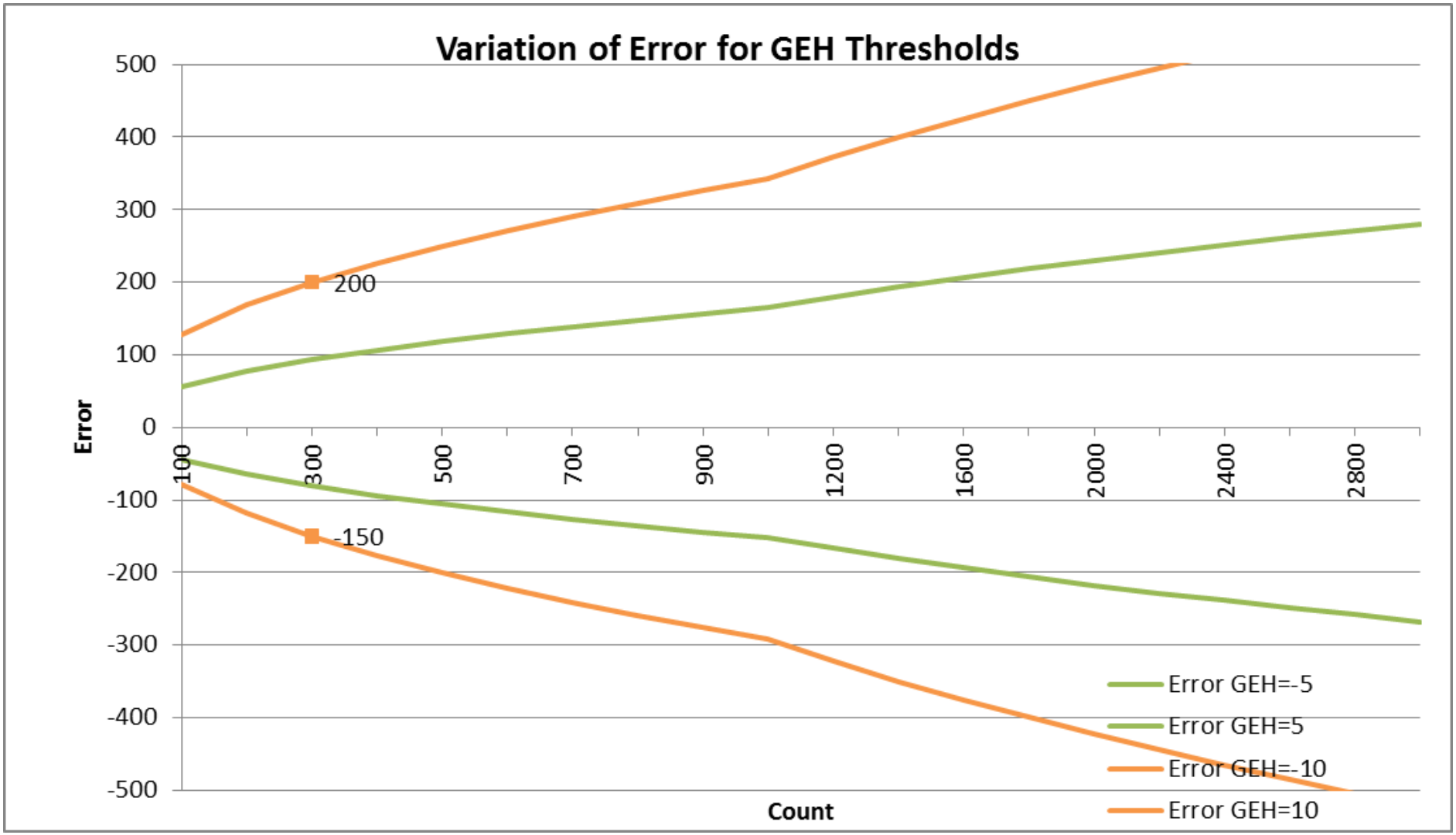
GEH Trend


Variation of GEH with Count for Constant Error
GEH
18
16
14
12
10
8
6
4
2
0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
Count
Error = 10%
Error = 20%
Error = 30%
Error = 40%

GEH Thresholds



GEH Thresholds



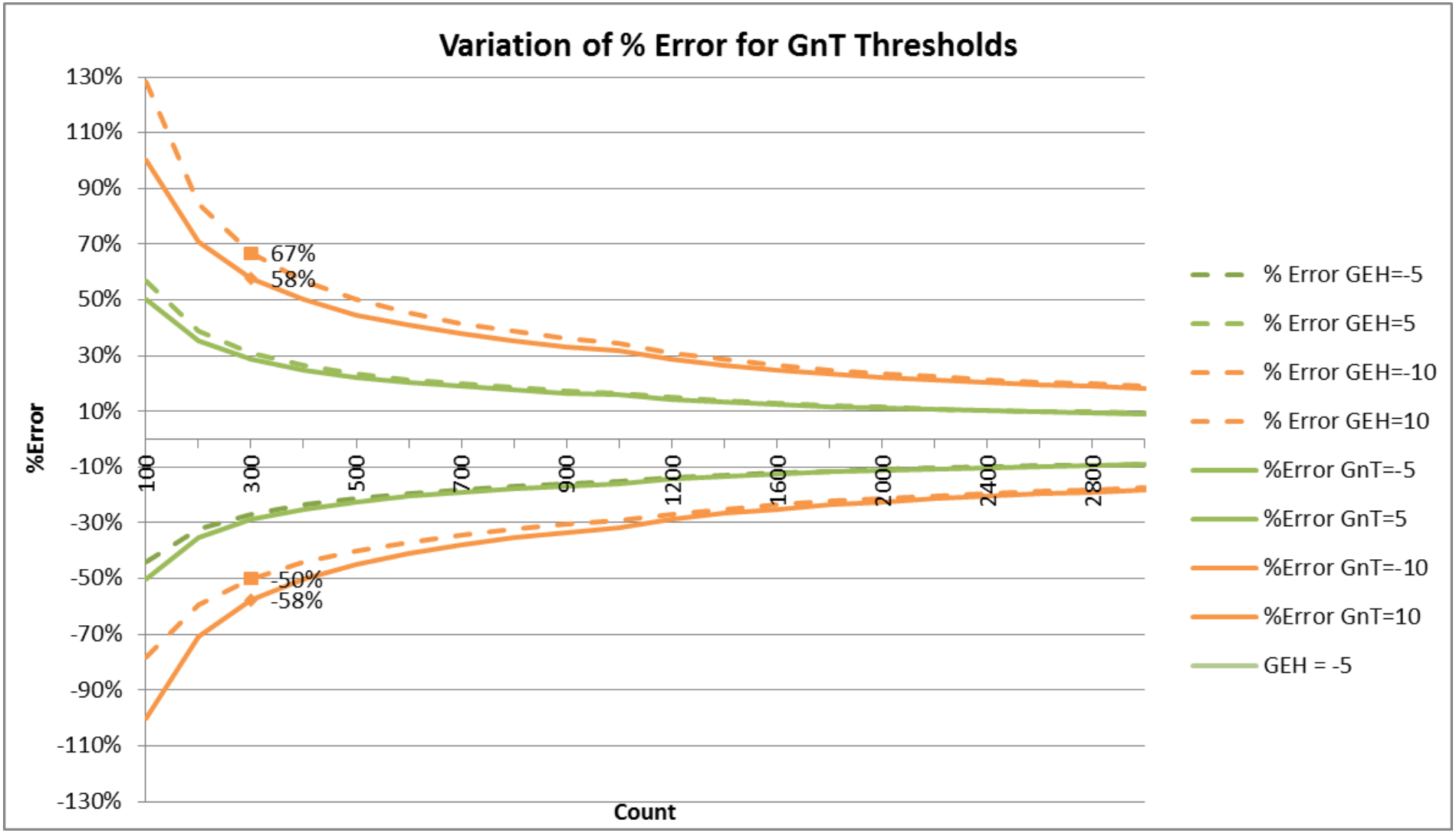
GEH Examples

	Model 1	Model 2
Count	400	400
Modelled	500	300
Error	100	-100
% Error	25%	-25%

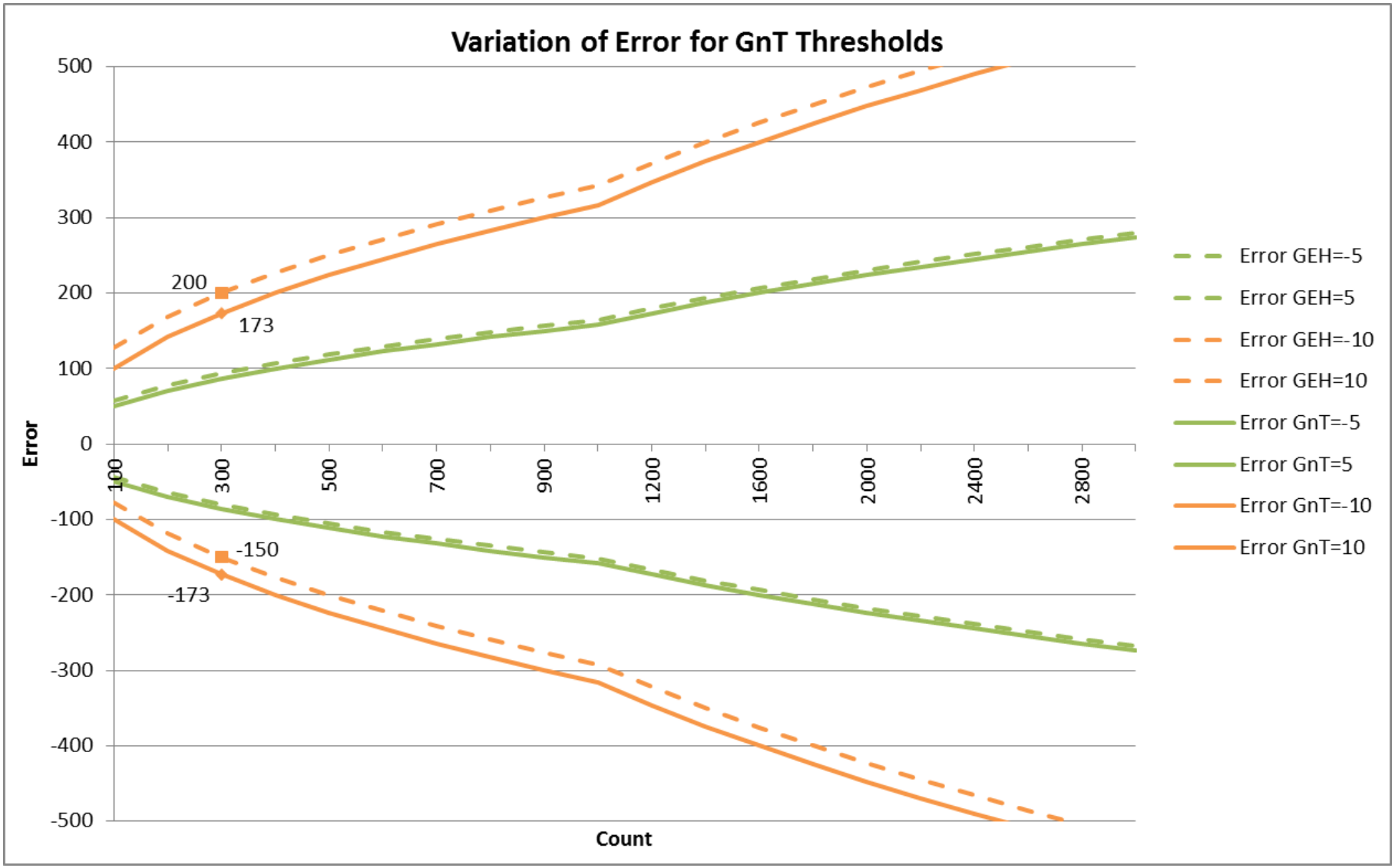
	Model 1	Model 2
Count	900	900
Modelled	600	1200
Error	-300	300
% Error	-33%	33%

- GEH: $\frac{Difference}{\sqrt{Average}}$
- $\sqrt{Average}$ Reflects uncertainty in the count
- What is the Purpose ?
  - To identify possible issues for further investigation
- GnT: $\frac{Difference}{\sqrt{Observed}}$

GnT Thresholds

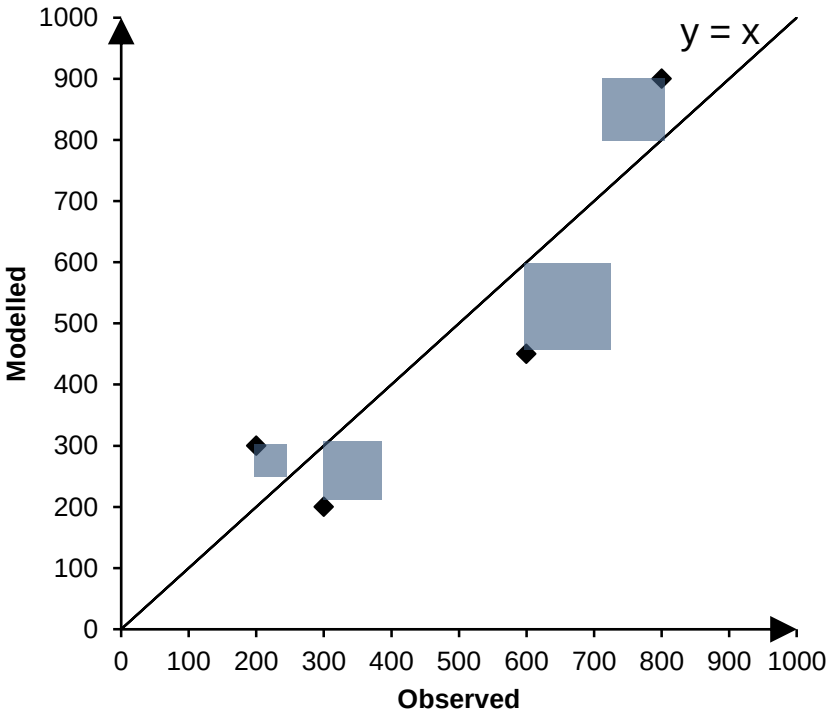


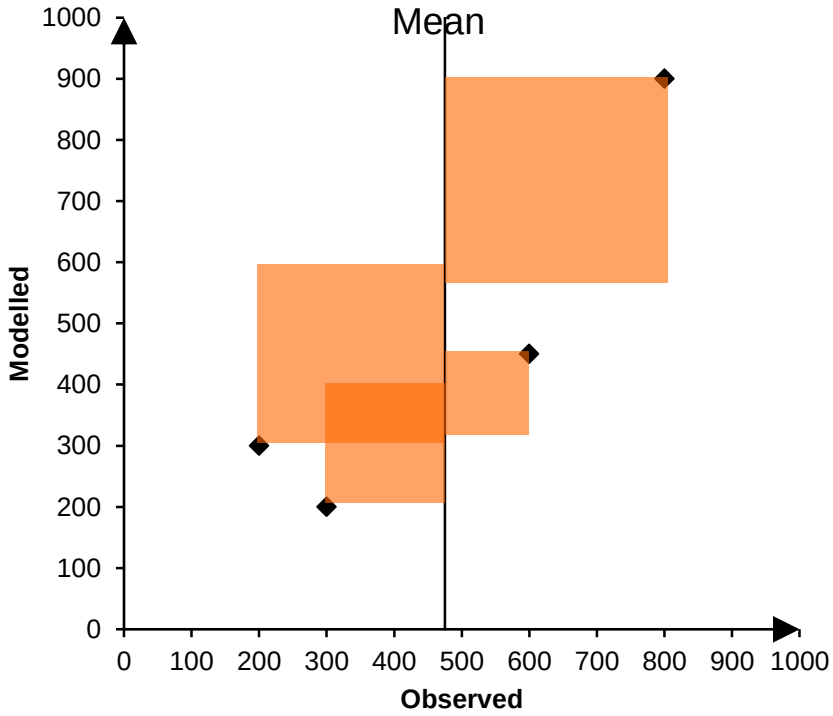
GnT Thresholds



R^2: Definition

- $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$
- $R^2 = 1 - \frac{Errors^2\ Summed}{Difference\ between\ a\ count\ and\ the\ mean\ count^2\ Summed}$





R^2: A Comparison

Observed	Modelled	GEH	% Error (M-O)/O	$Error^2$ $(M-O)^2$	$Diff^2$ $(O-Av)^2$
400	380	1.0	-5%		
450	473	1.0	+5%		
500	475	1.1	-5%		
505	530	1.1	+5%		
510	485	1.1	-5%		
515	541	1.1	+5%		
520	494	1.2	-5%		
525	551	1.1	+5%		
530	504	1.2	-5%		
600	630	1.2	+5%		

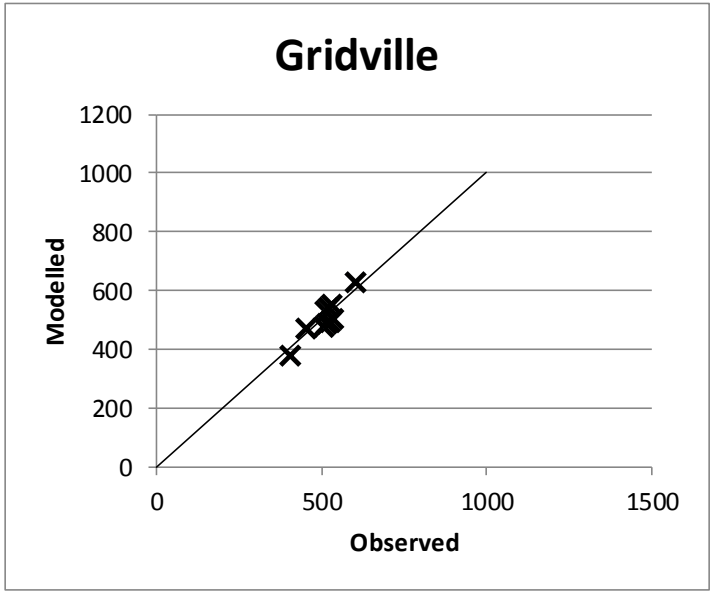
506 Ave **Totals:**

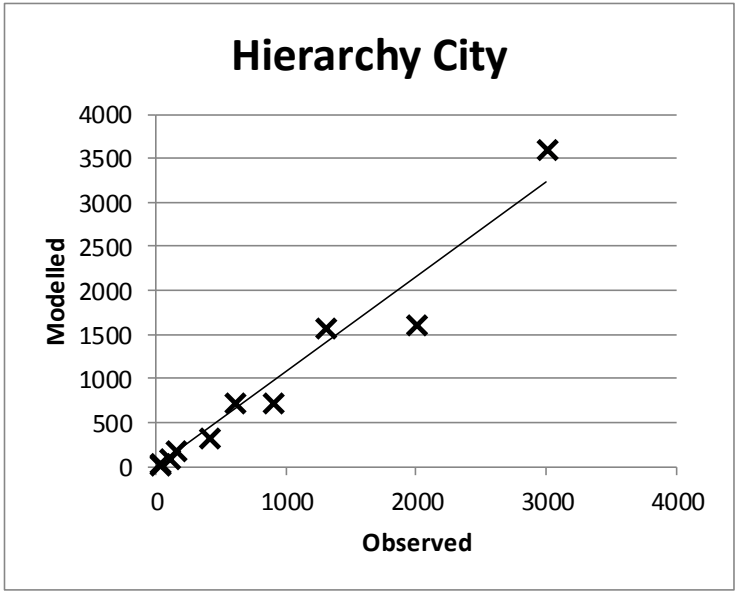
R^2:

Observed	Modelled	GEH	% Error (M-O)/O	$Error^2$ $(M-O)^2$	$Diff^2$ $(O-Av)^2$
20	16	0.9	-20%		
25	30	1.0	+20%		
100	80	2.1	-20%		
150	180	2.3	+20%		
400	320	4.2	-20%		
600	720	4.7	+20%		
900	720	6.3	-20%		
1300	1560	6.9	+20%		
2000	1600	9.4	-20%		
3000	3600	10.4	+20%		

850 Ave **Totals:**

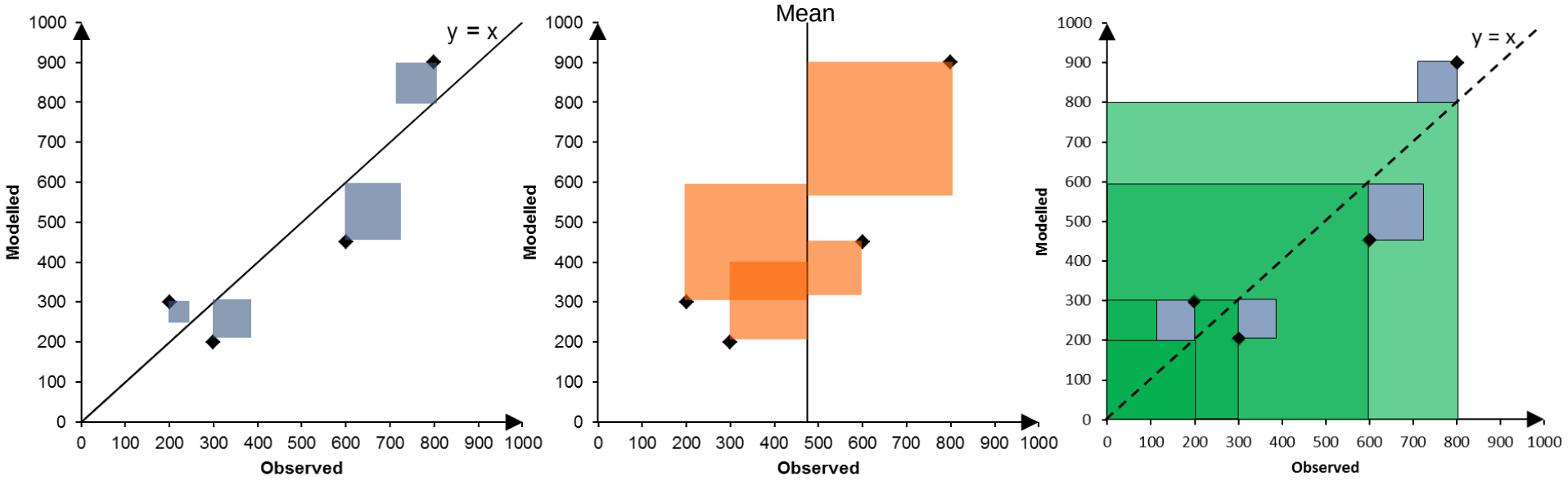
R^2:





R^2: An Alternative ?

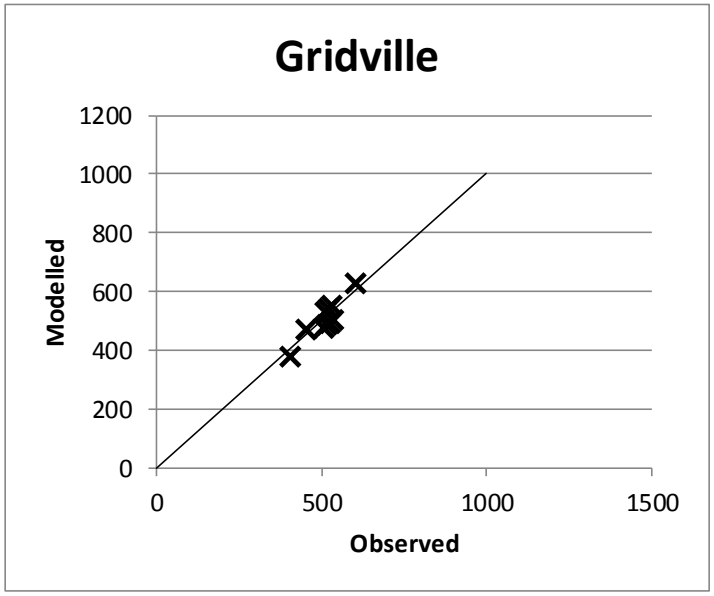
- $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$
- $R^2 = 1 - \frac{Error^2, Summed}{Difference\ between\ a\ Count\ and\ the\ Mean\ Count^2, Summed}$
- Alternative: $R^2 = 1 - \frac{Error^2, Summed}{Count^2, Summed}$

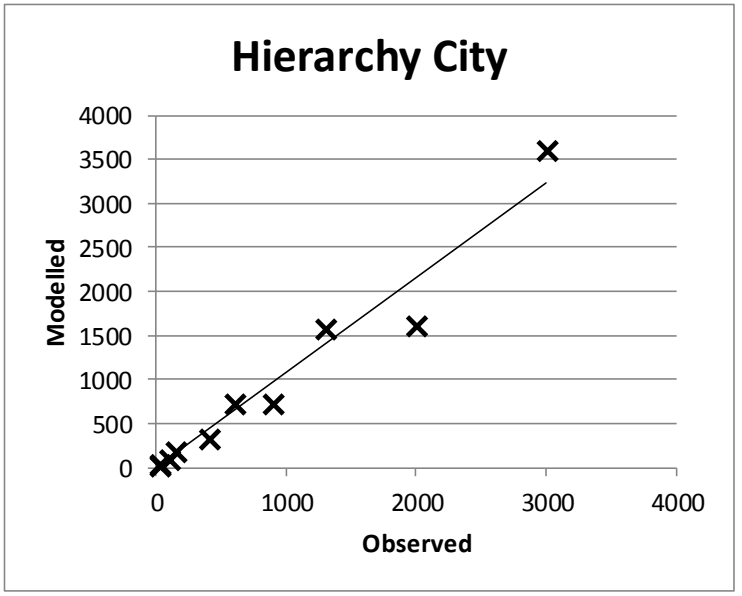


R^2: A Modification ?

Observed	Modelled	GEH	% Error (M-O)/O	Error^2 (M-O)^2	Obs^2 (O)^2
400	380	1.0	-5%	400	160,000
450	473	1.0	+5%	506	202,500
500	475	1.1	-5%	625	250,000
505	530	1.1	+5%	638	255,025
510	485	1.1	-5%	650	260,100
515	541	1.1	+5%	663	265,225
520	494	1.2	-5%	676	270,400
525	551	1.1	+5%	689	275,625
530	504	1.2	-5%	702	280,900
600	630	1.2	+5%	900	360,000
506 Ave			**Totals:**	6,449	2,579,775
			R^2:	**~~0.736~~**	**0.998**

Observed	Modelled	GEH	% Error (M-O)/O	Error^2 (M-O)^2	Obs^2 (O)^2
20	16	0.9	-20%	16	400
25	30	1.0	+20%	25	625
100	80	2.1	-20%	400	10,000
150	180	2.3	+20%	900	22,500
400	320	4.2	-20%	6,400	160,000
600	720	4.7	+20%	14,400	360,000
900	720	6.3	-20%	32,400	810,000
1300	1560	6.9	+20%	67,600	1,690,000
2000	1600	9.4	-20%	160,000	4,000,000
3000	3600	10.4	+20%	360,000	9,000,000
850 Ave			**Totals:**	642,141	16,053,525
			R^2:	**~~0.927~~**	**0.960**





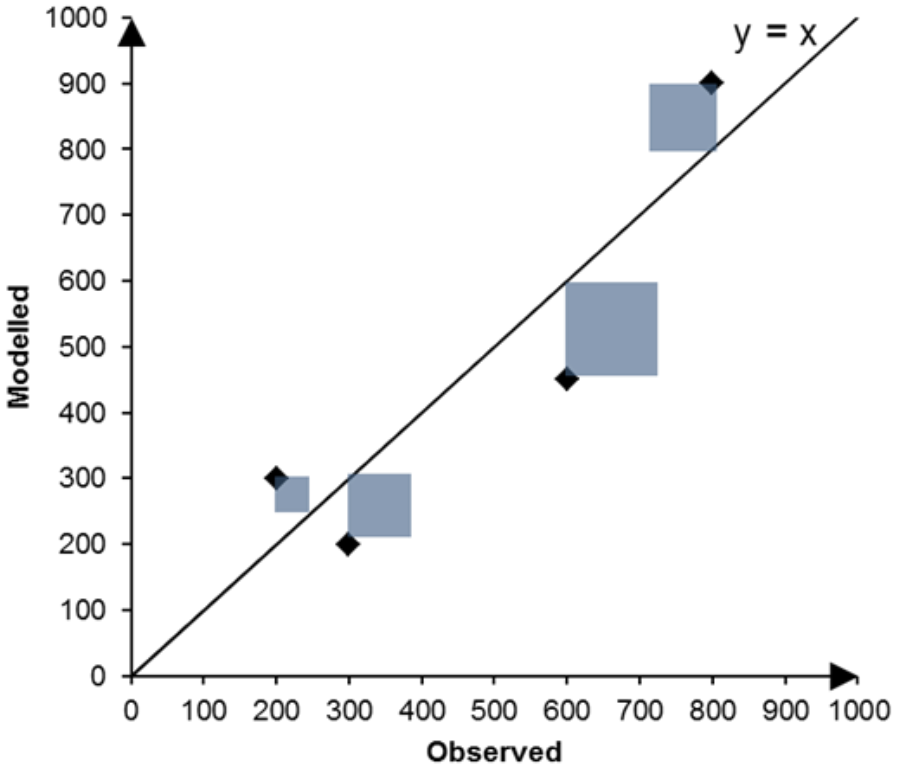
RMSE Definition (or %RMSE)

- % RMSE = $\frac{\sqrt{\frac{\sum(M-O)^2}{N-1}}}{\left[\frac{\sum O}{N}\right]}$ x 100

- $\frac{\sqrt{\frac{\sum Error^2}{N-1}}}{Ave\ Count}$

- $\frac{\sqrt{\frac{\sum Error^2}{N}}}{Ave\ Count}$

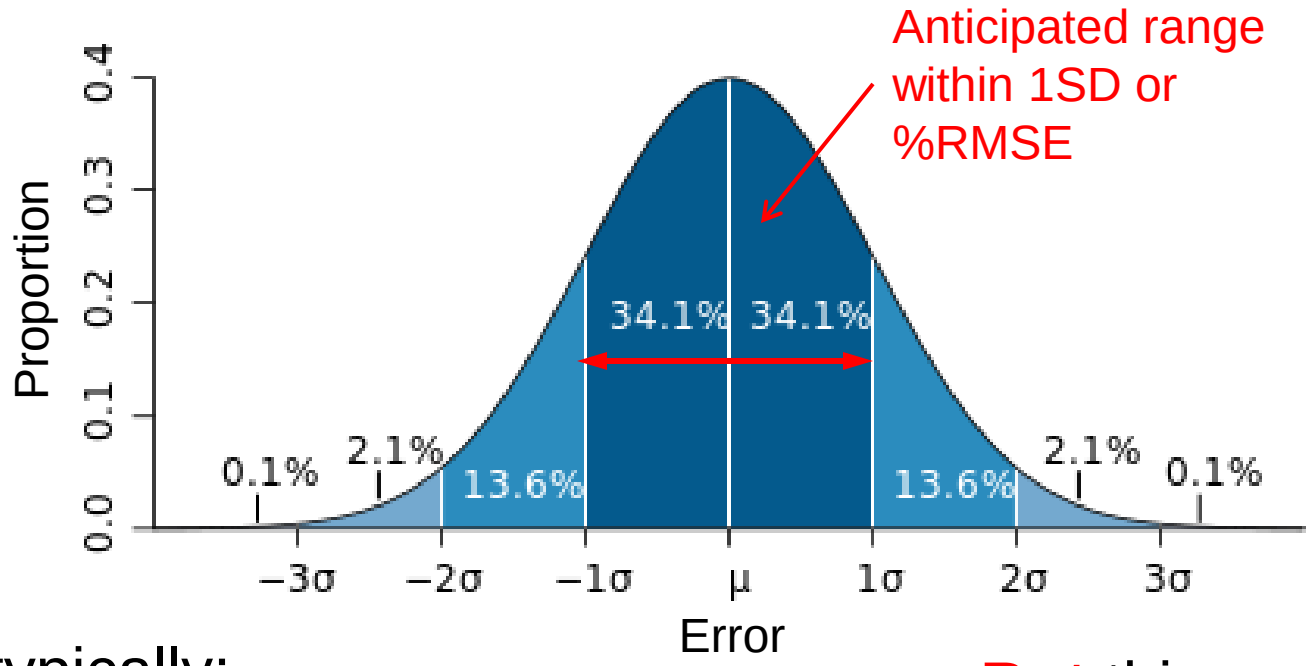
- $\frac{\sqrt{Average\ Squared\ Error}}{Ave\ Count}$



- What does an RMSE of 30% mean ?
- How is this a better measure than the 'weighted average % error', i.e:
  - $\frac{Average\ Absolute\ Difference}{Average\ Count}$?

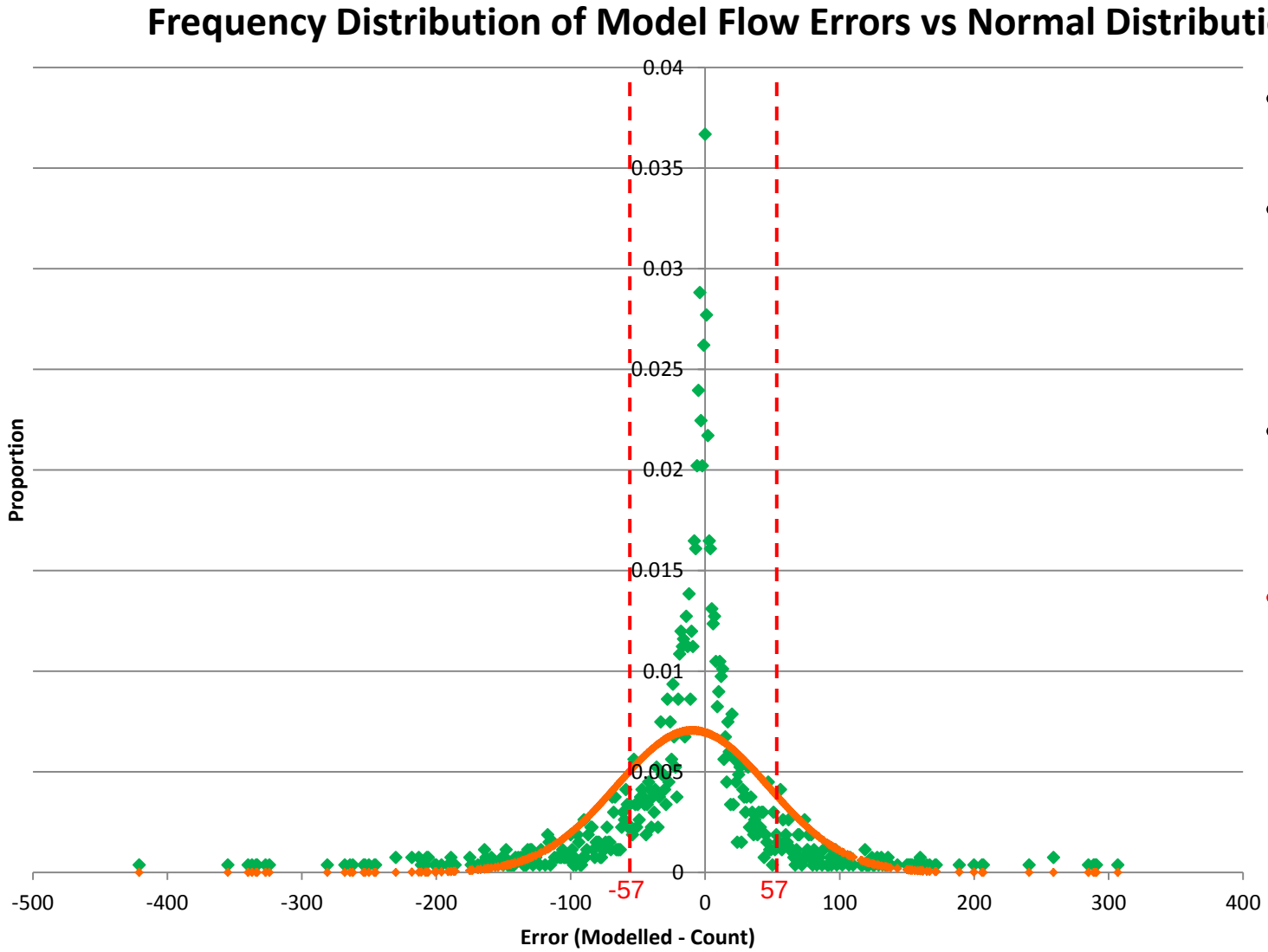
%RMSE As a Measure of Confidence

- RMSE closely related to Standard Deviation
- Theory is that SD (or RMSE) tells us something about the confidence in the model:



- i.e. typically:
  - 68% of errors < RMSE
  - 95% of errors < 2 x RMSE
- **But** this assumes a normal distribution of errors

%RMSE As a Measure of Confidence



- SD = 57 (RMSE = 24%)
- Normal Distribution Implies 68% of errors < SD
- Actual data shows 81% of data within SD
- So what is SD or RMSE actually telling us ?

%RMSE Thoughts

- Scatter plots are a better indication of overall model fit
- Is the %'Mean Absolute Deviation' i.e. $\frac{Average\ Absolute\ Difference}{Average\ Count}$ a more direct, simpler, intuitive single measure ?
- E.g. %MAD = 14% vs RMSE = 24%

Conclusions

- Is GnT a better indicator of potential issues with models than GEH ?
- Is R^2 appropriate to our purpose or should this be modified ?
- %RMSE not intuitive and of dubious value. Suggest replacing with %MAD
- Preference is to investigate & document reasons for all significant model vs. data discrepancies prior to and after any ME, rather than focussing on achieving a raft of arbitrary criteria.