Negative Binomial Regression Fit Report

20250610

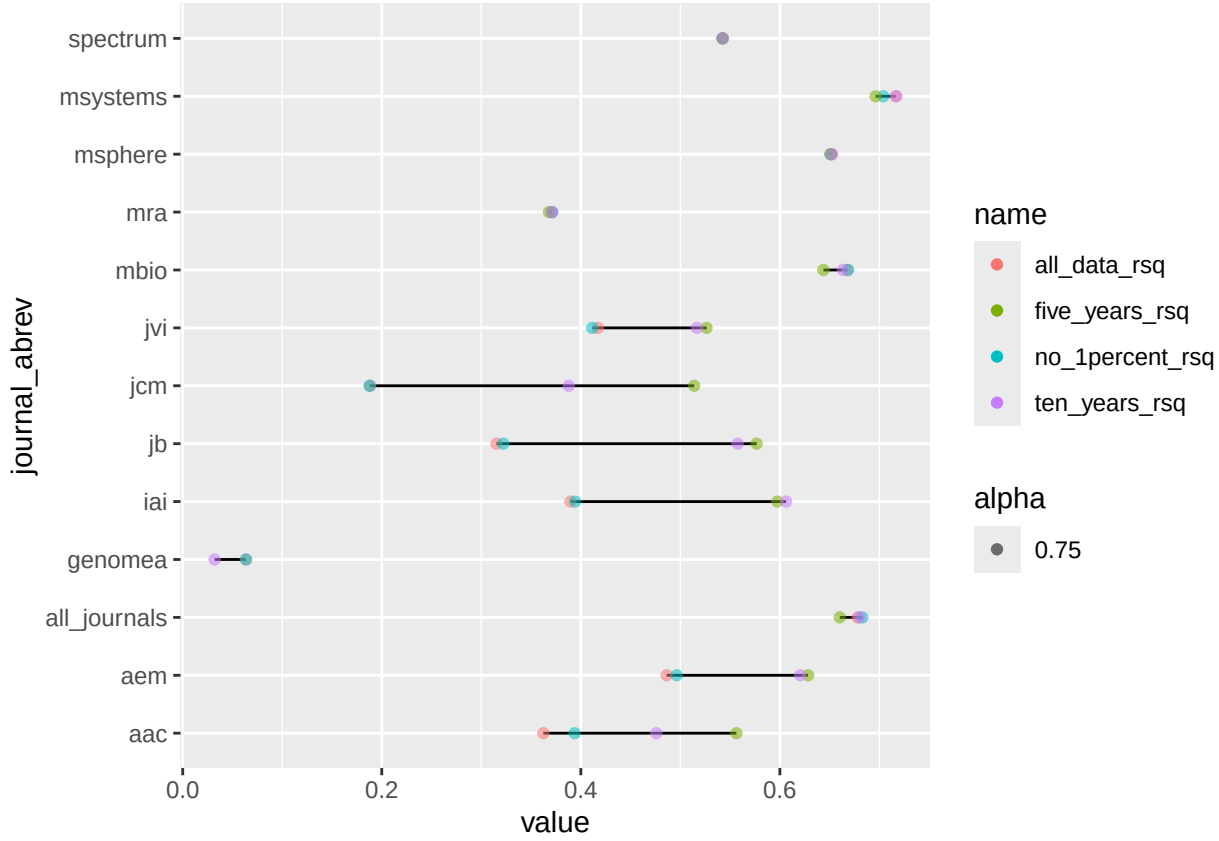
Project Summary

- We are using data from the American Society of Microbiology's (ASM) 12 published journals to investigate the relationship between the number of citations (variable 'is.referenced.by.count') a published scientific article receives and if the authors have included access to their raw sequencing data (variable 'da', data availability) in the manuscript.
- We are trying to understand if publishing raw data helps to improve citation metrics. We have data from 2000-2024, and will also adjust for time published (variable 'age.in.months'), as older papers have had the opportunity to accumulate more citations over time.

```
knitr::kable(rsquared, digits = 4)
```

journal_abrev	n	all_data_rsq	no_1percent_rsq	five_years_rsq	ten_years_rsq
aac	3237	0.3623	0.3938	0.5564	0.4759
aem	8638	0.4862	0.4964	0.6284	0.6204
genomea	6578	0.0636	0.0636	NA	0.0321
iai	1854	0.3896	0.3943	0.5975	0.6062
jb	4867	0.3152	0.3221	0.5767	0.5575
jcm	4374	0.1882	0.1878	0.5139	0.3880
jvi	4583	0.4172	0.4115	0.5264	0.5167
mbio	2498	0.6680	0.6685	0.6438	0.6633
mra	5738	0.3712	0.3712	0.3680	0.3712
msphere	1041	0.6523	0.6508	0.6510	0.6523
msystems	1436	0.7168	0.7040	0.6962	0.7168
spectrum	2957	0.5425	0.5425	0.5425	0.5425
all_journals	47808	0.6781	0.6829	0.6602	0.6801

```
rsquared %>%
  pivot_longer(cols = all_data_rsq:ten_years_rsq) %>%
  ggplot(aes(y = journal_abrev, x = value)) +
  geom_line(na.rm = TRUE) +
  geom_point(aes(color = name, alpha = 0.75), na.rm = TRUE)
```



How well do the models fit (by Cragg-Uhler pseduo R-squared metric)

- See above table "rsquared"

- Model format for all data from all journals MASS::glm.nb(is.referenced.by.count~ da_factor + log(age.in.months) + container.title + container.title*da_factor + log(age.in.months)*da_factor + container.title*log(age.in.months) + log(age.in.months)*da_factor*container.title, data = nsd_yes_metadata, link = log)
- Use model format for data from each journal MASS::glm.nb(is.referenced.by.count~ da_factor + log(age.in.months) + log(age.in.months)*da_factor, data = <each journal>, link = log)
- **Overall model fit with all data from all journals:**
  - R^2 value = 0.678
  - Removal of top 1% of data: R^2 value = 0.682
  - Truncate data to last 5 years: R^2 value = 0.660
  - Truncate data to last 10 years: R^2 value = 0.680
  - **Summary :** Model fit by R^2 metric does not change by removing the top 1% of data or truncating to data from the last 5 or 10 years.
- **Overall model fit for data from EACH journal individually:**
  - 4/12 journals have **overall model fit** with R^2 > 0.5
  - 4/12 journals have fit with R^2 > 0.5 with **top 1% of data removed**
  - 10/11 journals have model fits >0.5 when **truncated to the last 5 years,** so they are better than their fit overall (one journal has no data from this period)
  - 8/12 journals have model fits >0.5 when **truncated to the last 10 years,** so they are better than their fit overall
  - **Summary:** Data fits negative binomial model better with only more recent data considered.

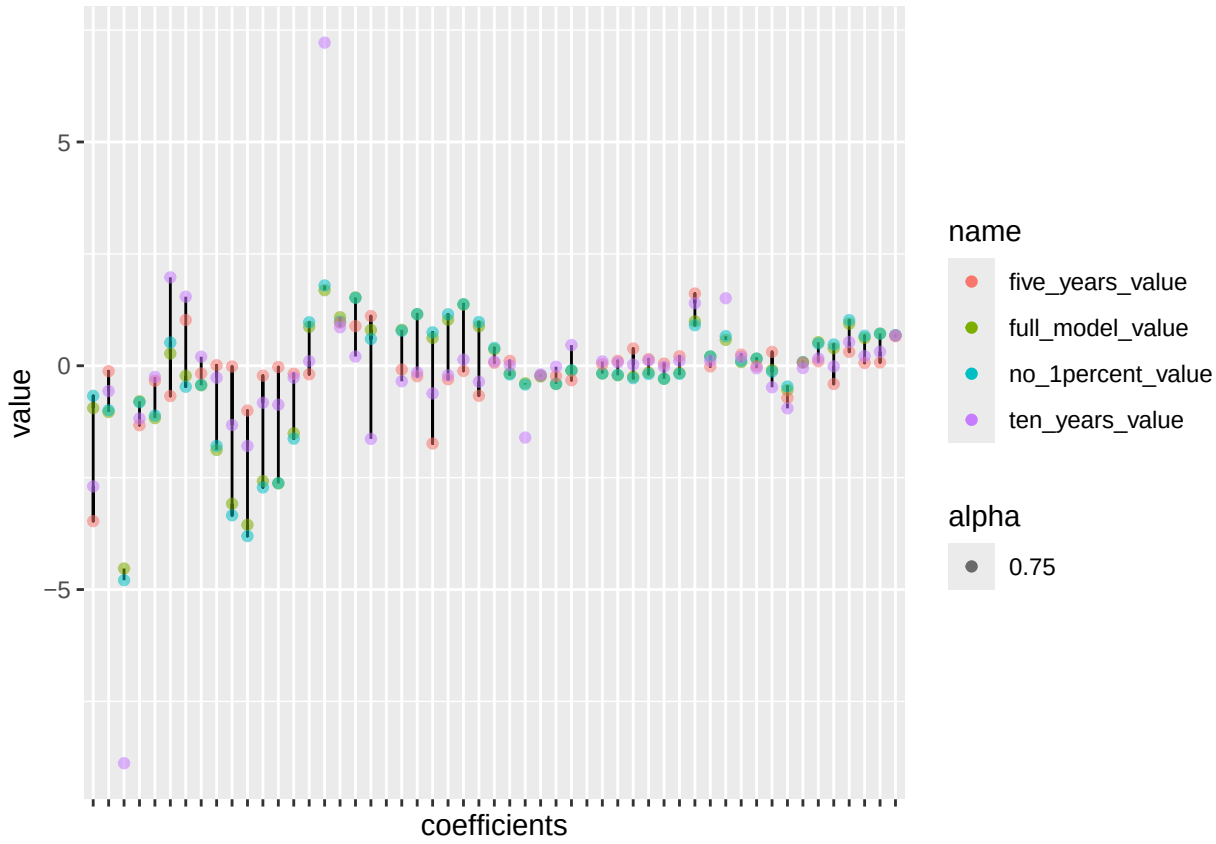
```
knitr::kable(all_journals, digits = 4)
```

coefficients	full_model_value	no_1percent_value	five_years_value	ten_years_value
rsquared	0.6781	0.6829	0.6602	0.6801
(Intercept)	-0.9460	-0.6722	-3.4726	-2.6960
da_factorYes	-1.5082	-1.6296	-0.1780	-0.2667
log(age.in.months)	0.9932	0.9069	1.6141	1.3976
container.titleApplied and Environmental Microbiology	-1.0337	-0.9933	-0.1207	-0.5651
container.titleGenome Announcements	-4.5303	-4.7923	NA	-8.8809
container.titleInfection and Immunity	-0.7927	-0.8150	-1.3290	-1.1797
container.titleJournal of Bacteriology	-1.1706	-1.1146	-0.3361	-0.2502
container.titleJournal of Clinical Microbiology	0.2729	0.5180	-0.6776	1.9759
container.titleJournal of Microbiology & Biology Education	-0.2261	-0.4692	1.0220	1.5424

coefficients	full_model_value	no_1percent_value	five_years_value	ten_years_value
container.titleJournal of Virology	-0.4370	-0.4275	-0.1728	0.2013
container.titlemBio	-1.8817	-1.7910	0.0115	-0.2660
container.titleMicrobiology Resource Announcements	-3.0840	-3.3426	-0.0170	-1.3247
container.titleMicrobiology Spectrum	-3.5550	-3.8058	-1.0015	-1.7913
container.titlemSphere	-2.5767	-2.7184	-0.2216	-0.8267
container.titlemSystems	-2.6293	-2.6203	-0.0287	-0.8701
da_factorYes:container.titleApplied and Environmental Microbiology	0.8693	0.9714	-0.1907	0.1001
da_factorYes:container.titleGenome Announcements	1.6881	1.7938	NA	7.2193
da_factorYes:container.titleInfection and Immunity	1.0848	0.9853	0.9639	0.8601
da_factorYes:container.titleJournal of Bacteriology	1.5260	1.5178	0.8850	0.2051
da_factorYes:container.titleJournal of Clinical Microbiology	0.7994	0.6000	1.1103	-1.6342
da_factorYes:container.titleJournal of Microbiology & Biology Education	NA	NA	NA	NA
da_factorYes:container.titleJournal of Virology	0.8019	0.7830	-0.0860	-0.3469
da_factorYes:container.titlemBio	1.1512	1.1542	-0.2339	-0.1494
da_factorYes:container.titleMicrobiology Resource Announcements	0.6208	0.7466	-1.7384	-0.6183
da_factorYes:container.titleMicrobiology Spectrum	1.0273	1.1494	-0.3022	-0.2138
da_factorYes:container.titlemSphere	1.3657	1.3770	-0.1163	0.1374
da_factorYes:container.titlemSystems	0.8815	0.9763	-0.6710	-0.3577
da_factorYes:log(age.in.months)	0.3675	0.3986	0.0661	0.0957
log(age.in.months):container.titleApplied and Environmental Microbiology	0.2058	0.2073	-0.0156	0.1092
log(age.in.months):container.titleGenome Announcements	0.5778	0.6617	NA	1.5084
log(age.in.months):container.titleInfection and Immunity	0.0813	0.1089	0.2478	0.1951
log(age.in.months):container.titleJournal of Bacteriology	0.1570	0.1612	-0.0098	-0.0566
log(age.in.months):container.titleJournal of Clinical Microbiology	-0.0931	-0.1295	0.3100	-0.4809
log(age.in.months):container.titleJournal of Microbiology & Biology Education	-0.5427	-0.4649	-0.7173	-0.9522
log(age.in.months):container.titleJournal of Virology	0.0693	0.0849	0.0714	-0.0526
log(age.in.months):container.titlemBio	0.5267	0.4983	0.0984	0.1610
log(age.in.months):container.titleMicrobiology Resource Announcements	0.3927	0.4751	-0.4083	-0.0141
log(age.in.months):container.titleMicrobiology Spectrum	0.9389	1.0180	0.3094	0.5302
log(age.in.months):container.titlemSphere	0.6216	0.6690	0.0632	0.2172
log(age.in.months):container.titlemSystems	0.7187	0.7154	0.0735	0.3119
da_factorYes:log(age.in.months):container.titleApplied and Environmental Microbiology	-0.1698	-0.1948	0.1092	0.0088
da_factorYes:log(age.in.months):container.titleGenome Announcements	-0.3920	-0.4198	NA	-1.6043
da_factorYes:log(age.in.months):container.titleInfection and Immunity	-0.2392	-0.2182	-0.2194	-0.1933
da_factorYes:log(age.in.months):container.titleJournal of Bacteriology	-0.4067	-0.4056	-0.2304	-0.0223
da_factorYes:log(age.in.months):container.titleJournal of Clinical Microbiology	-0.1115	-0.0999	-0.3246	0.4618

coefficients	full_model_value	no_1percent_value	five_years_value	ten_years_value
da_factorYes:log(age.in.months):container.titleJournal of Microbiology & Biology Education	NA	NA	NA	NA
da_factorYes:log(age.in.months):container.titleJournal of Virology	-0.1740	-0.1730	0.0329	0.0975
da_factorYes:log(age.in.months):container.titlemBio	-0.2141	-0.2046	0.1095	0.0758
da_factorYes:log(age.in.months):container.titleMicrobiology Resource Announcements	-0.2456	-0.2778	0.3859	0.0256
da_factorYes:log(age.in.months):container.titleMicrobiology Spectrum	-0.1533	-0.1846	0.1478	0.1183
da_factorYes:log(age.in.months):container.titlemSphere	-0.2926	-0.2903	0.0498	-0.0242
da_factorYes:log(age.in.months):container.titlemSystems	-0.1622	-0.1815	0.2131	0.1090

```
all_journals %>%
  pivot_longer(cols = full_model_value:ten_years_value) %>%
  ggplot(aes(x = coefficients, y = value)) +
  geom_line(na.rm = TRUE) +
  geom_point(aes(color = name, alpha = 0.75), na.rm = TRUE) + theme(axis.text.x = element_blank())
```



All journal model is resistant to changes from removing top 1% of data, but less resistant to changes from truncating at 5 and 10 years.

- When working across the columns in the second table, we have coefficients on the left, followed by their values under the following conditions
  - full_model_value = all data included in the model
  - no_1percent_value = top 1% of data removed
  - five_years_value = data truncated at 5 years in age of paper
  - ten_years_value = data truncated at 10 years in age of paper

  - **Note:** Journal of Microbiology and Biology Education(jmbe) has N=7 papers with new sequence data and has been excluded for these analyses, but is a part of the model, and appears as NAs in the table above.

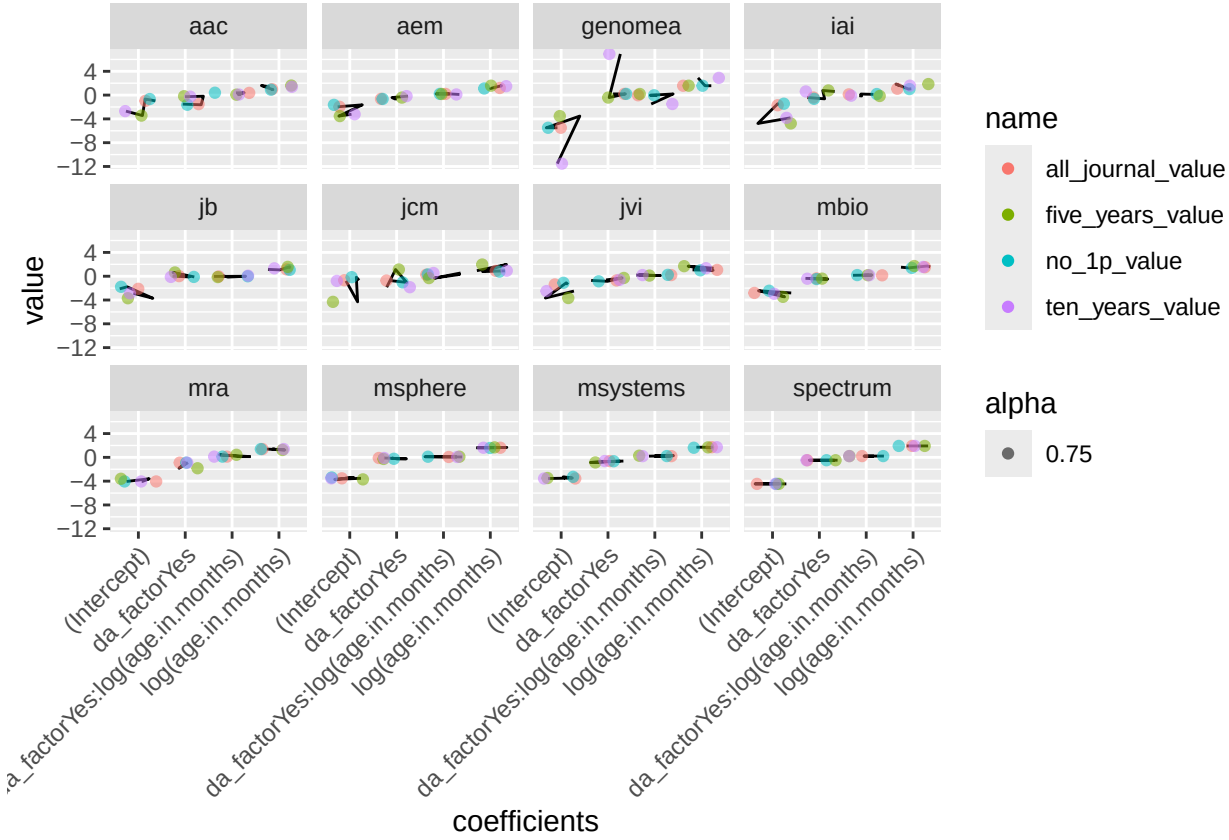
- Is this graphic helpful? The labels are super cluttered and need some work.

```
knitr::kable(each_journal, digits = 4)
```

name	coefficients	all_journal_value	no_1p_value	five_years_value	ten_years_value
aac	(Intercept)	-0.9609	-0.6665	-3.4423	-2.6960
aac	da_factorYes	-1.5204	-1.6244	-0.1890	-0.2667
aac	log(age.in.months)	0.9961	0.9057	1.6058	1.3976
aac	da_factorYes:log(age.in.months)	0.3708	0.3972	0.0691	0.0957
aem	(Intercept)	-1.9654	-1.6400	-3.5278	-3.2253
aem	da_factorYes	-0.6343	-0.6462	-0.4172	-0.1718
aem	log(age.in.months)	1.1963	1.1092	1.5802	1.4985
aem	da_factorYes:log(age.in.months)	0.1963	0.2004	0.1888	0.1054
genomea	(Intercept)	-5.4902	-5.4902	-3.5278	-11.5225
genomea	da_factorYes	0.1984	0.1984	-0.4172	6.9105
genomea	log(age.in.months)	1.5740	1.5740	1.5802	2.8942
genomea	da_factorYes:log(age.in.months)	-0.0284	-0.0284	0.1888	-1.4995
iai	(Intercept)	-1.7012	-1.4506	-4.7673	-3.8227
iai	da_factorYes	-0.4050	-0.6297	0.7774	0.6081
iai	log(age.in.months)	1.0675	1.0090	1.8523	1.5803
iai	da_factorYes:log(age.in.months)	0.1231	0.1761	-0.1510	-0.1015
jb	(Intercept)	-2.1151	-1.7806	-3.7023	-2.8669
jb	da_factorYes	0.0163	-0.1174	0.6339	-0.1168
jb	log(age.in.months)	1.1499	1.0669	1.5749	1.3225
jb	da_factorYes:log(age.in.months)	-0.0389	-0.0060	-0.1444	0.0859
jcm	(Intercept)	-0.6815	-0.1575	-4.3089	-0.7888
jcm	da_factorYes	-0.7281	-1.0364	1.1310	-1.8444
jcm	log(age.in.months)	0.9018	0.7780	1.9686	0.9325
jcm	da_factorYes:log(age.in.months)	0.2605	0.3003	-0.3140	0.5445
jvi	(Intercept)	-1.3855	-1.1081	-3.6656	-2.4940
jvi	da_factorYes	-0.7088	-0.8540	-0.2909	-0.6135
jvi	log(age.in.months)	1.0630	0.9935	1.6911	1.3449
jvi	da_factorYes:log(age.in.months)	0.1941	0.2275	0.1068	0.1932
mbio	(Intercept)	-2.8112	-2.4410	-3.4651	-2.9554

name	coefficients	all_journal_value	no_1p_value	five_years_value	ten_years_value
mbio	da_factorYes	-0.3597	-0.4783	-0.4089	-0.4184
mbio	log(age.in.months)	1.5161	1.4000	1.7137	1.5569
mbio	da_factorYes:log(age.in.months)	0.1539	0.1944	0.1747	0.1720
mra	(Intercept)	-4.0497	-4.0497	-3.5995	-4.0497
mra	da_factorYes	-0.8906	-0.8906	-1.8301	-0.8906
mra	log(age.in.months)	1.3911	1.3911	1.2378	1.3911
mra	da_factorYes:log(age.in.months)	0.1228	0.1228	0.4267	0.1228
msphere	(Intercept)	-3.5233	-3.3874	-3.7000	-3.5233
msphere	da_factorYes	-0.1065	-0.2401	-0.2685	-0.1065
msphere	log(age.in.months)	1.6149	1.5751	1.6790	1.6149
msphere	da_factorYes:log(age.in.months)	0.0657	0.1050	0.1088	0.0657
msystems	(Intercept)	-3.5575	-3.2742	-3.4729	-3.5575
msystems	da_factorYes	-0.6220	-0.6507	-0.8537	-0.6220
msystems	log(age.in.months)	1.7072	1.6174	1.6794	1.7072
msystems	da_factorYes:log(age.in.months)	0.2040	0.2163	0.2807	0.2040
spectrum	(Intercept)	-4.4608	-4.4608	-4.4608	-4.4608
spectrum	da_factorYes	-0.4802	-0.4802	-0.4802	-0.4802
spectrum	log(age.in.months)	1.9194	1.9194	1.9194	1.9194
spectrum	da_factorYes:log(age.in.months)	0.2139	0.2139	0.2139	0.2139

```
each_journal %>%
    rename(journal_abrev = name) %>%
  pivot_longer(cols = all_journal_value:ten_years_value) %>%
  ggplot(aes(x = coefficients, y = value)) +
  geom_line(na.rm = TRUE, position = "jitter") +
  geom_point(aes(color = name, alpha = 0.75), na.rm = TRUE, position = "jitter") +
  facet_wrap(vars(journal_abrev)) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```



Each journal model are semi-resistant to changes from removing top 1% of data, and even less resistant to changes from truncating at 5 and 10 years.

- See above for mutations on these columns, but these models look less resistant to the transformation of removing the top 1% of data, and even less resistant to changes in coefficients from truncating at 5 and 10 years of data.
- Is this graphic helpful? I think the scale should be adjusted to better show the data.