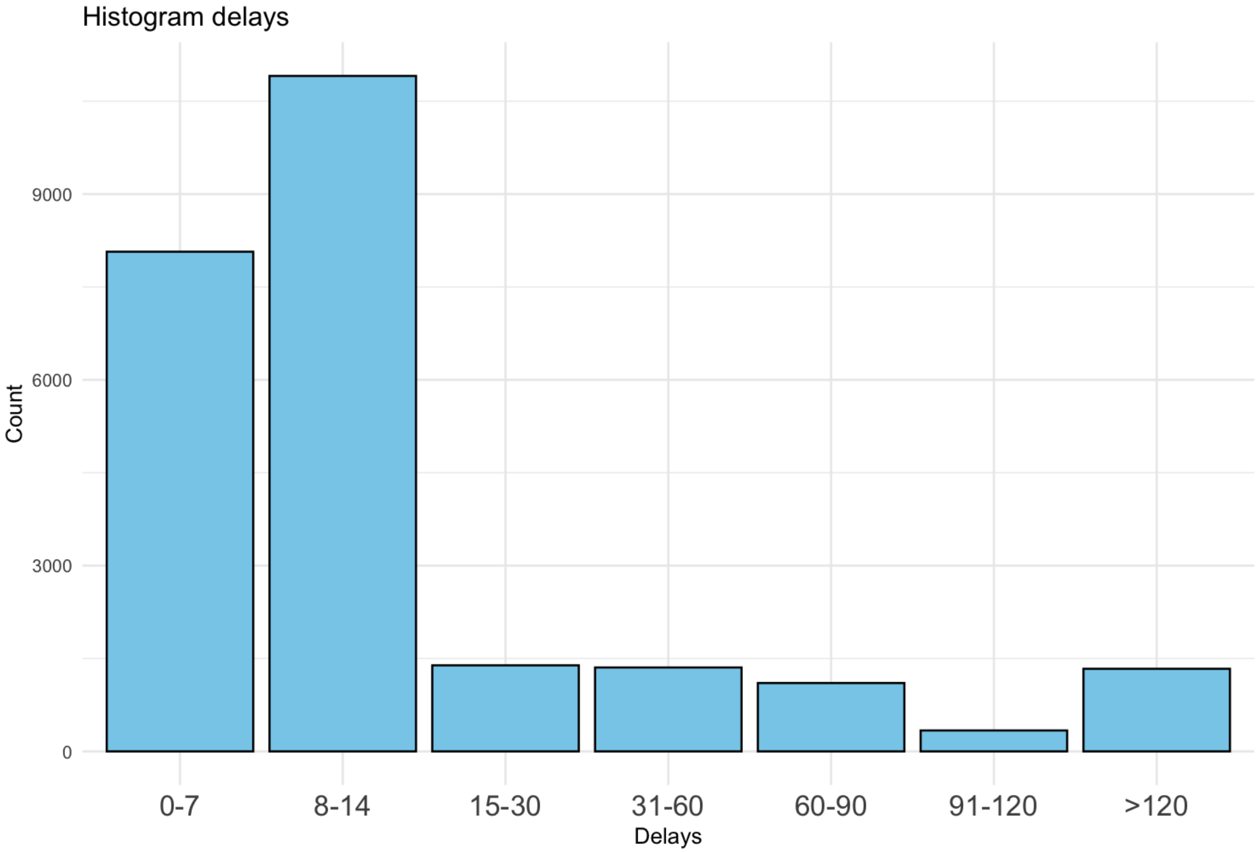
Presentation

GLOBAL VIEW : DESCRIPTIVE

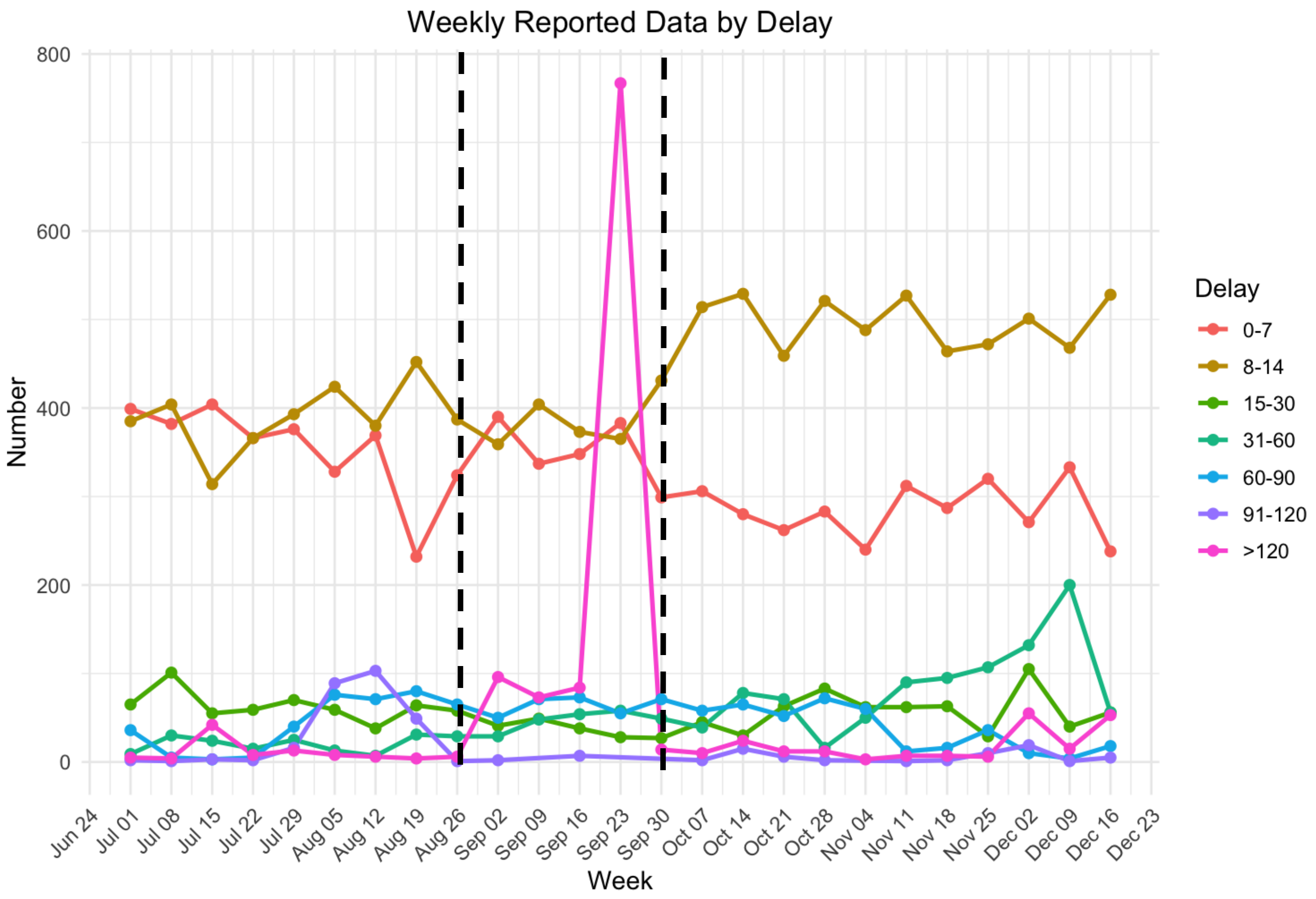


Data: the events that newly appeared after the first download are assumed to be newly reported

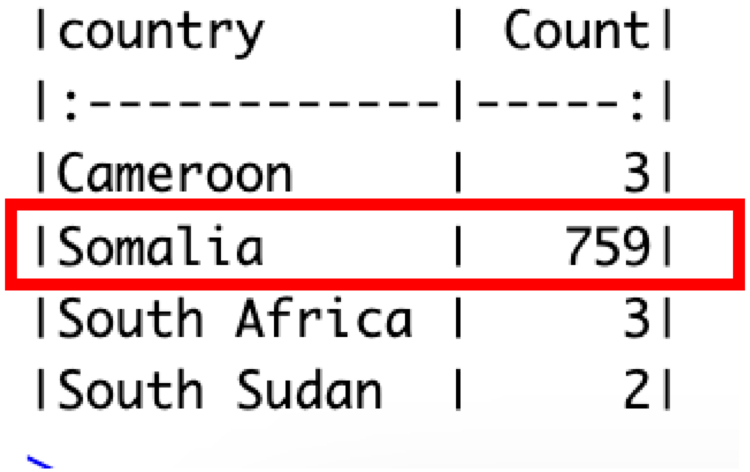
Delay: date of first report - date of event

- The date of report is subject to change when there is an update (sometimes an error correction)
- For the global view of the delay we consider the first date of report and omit error correction (this comes later)

GLOBAL VIEW : DESCRIPTIVE



Number of events reported weekly by delay



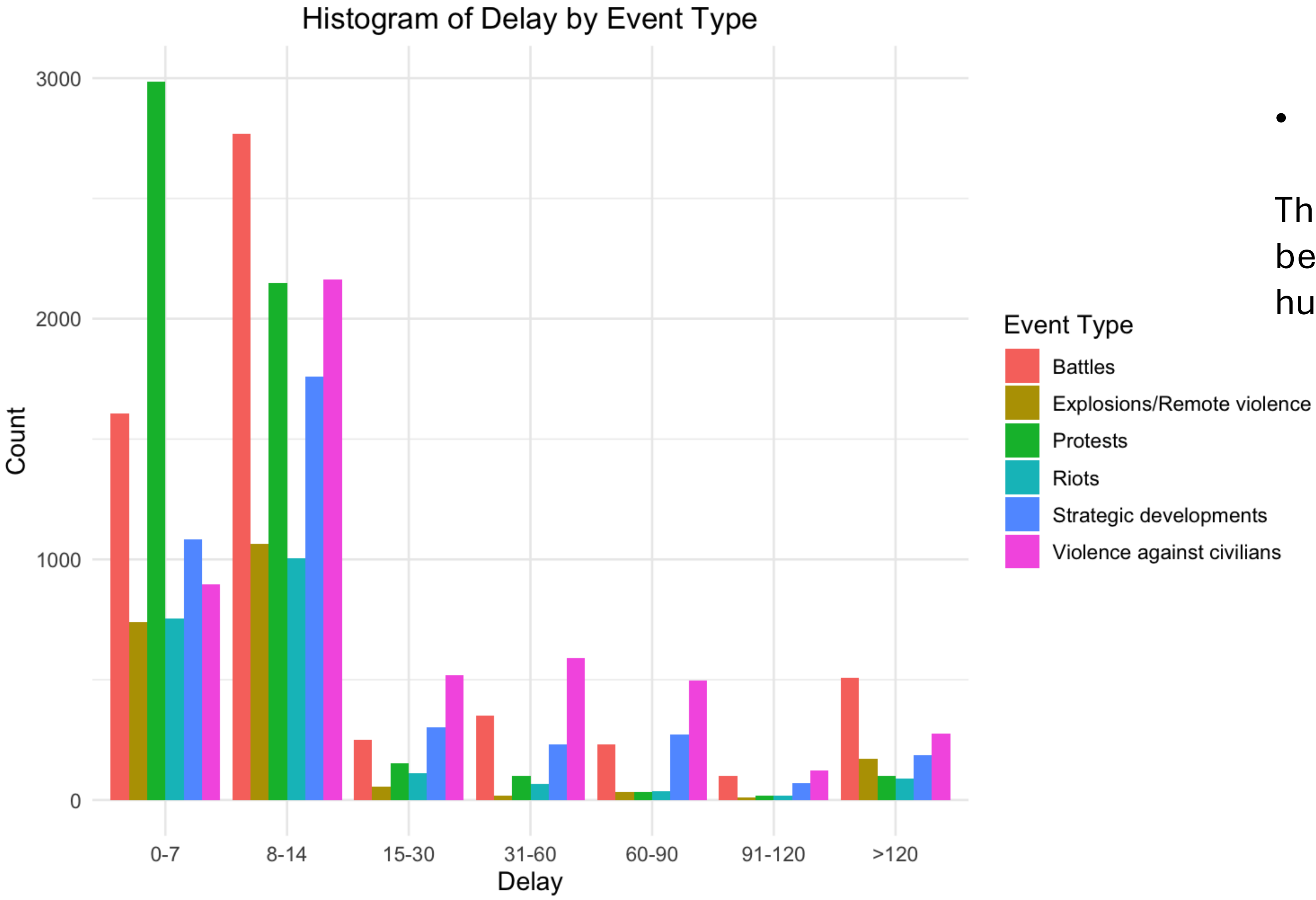
```
|country      | Count|
|:------------|-----:|
|Cameroon     |     3|
|Somalia      |   759|
|South Africa |     3|
|South Sudan  |     2|
```

We have sko symmetry between 0-7 and 8-14 (this indicates a negative correlation)

Corr = -0.8

Are the delays autocorrelated? (and of what order)

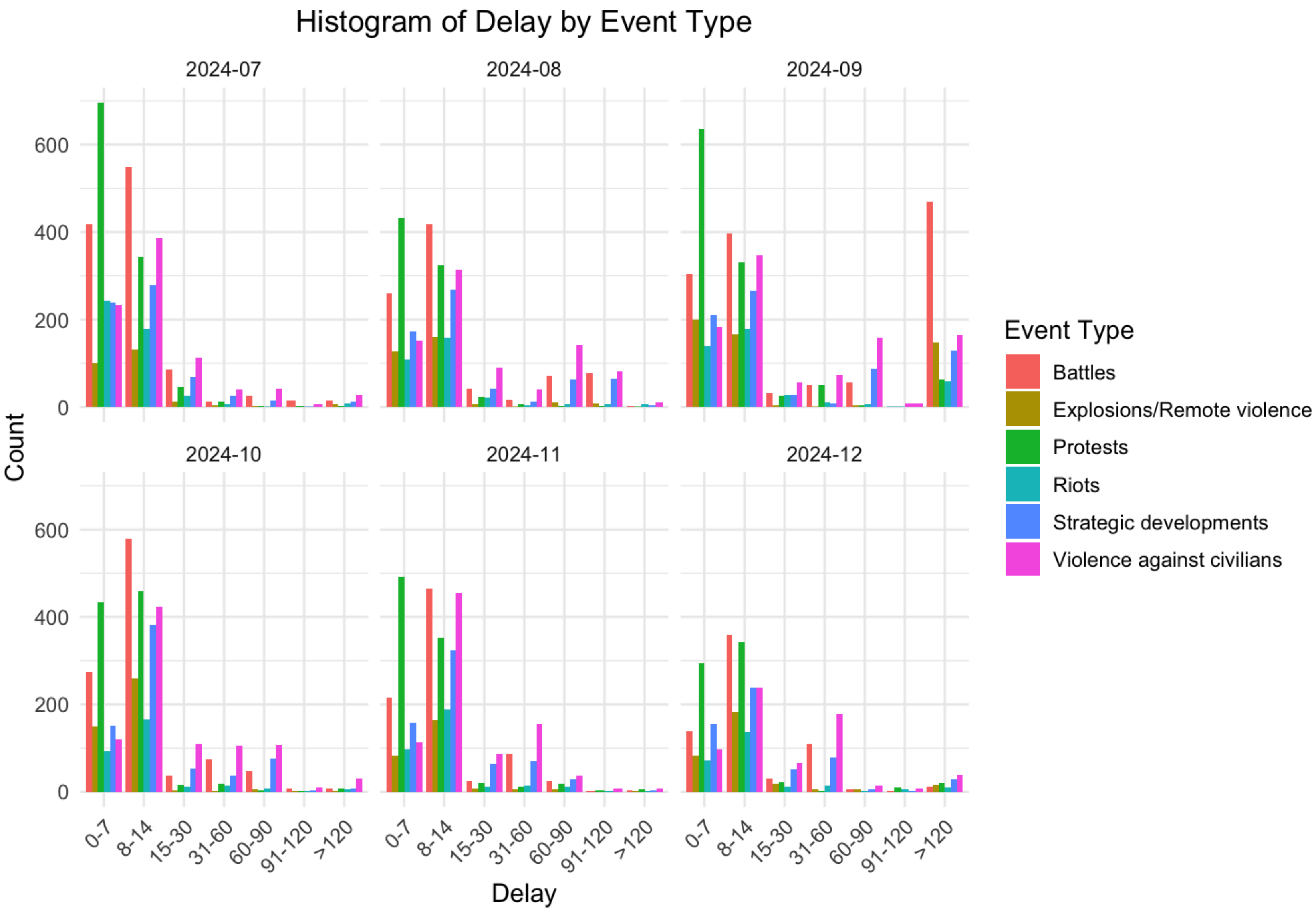
GLOBAL VIEW : DESCRIPTIVE



- How delays are distributed between event types?
- Does it change over time?

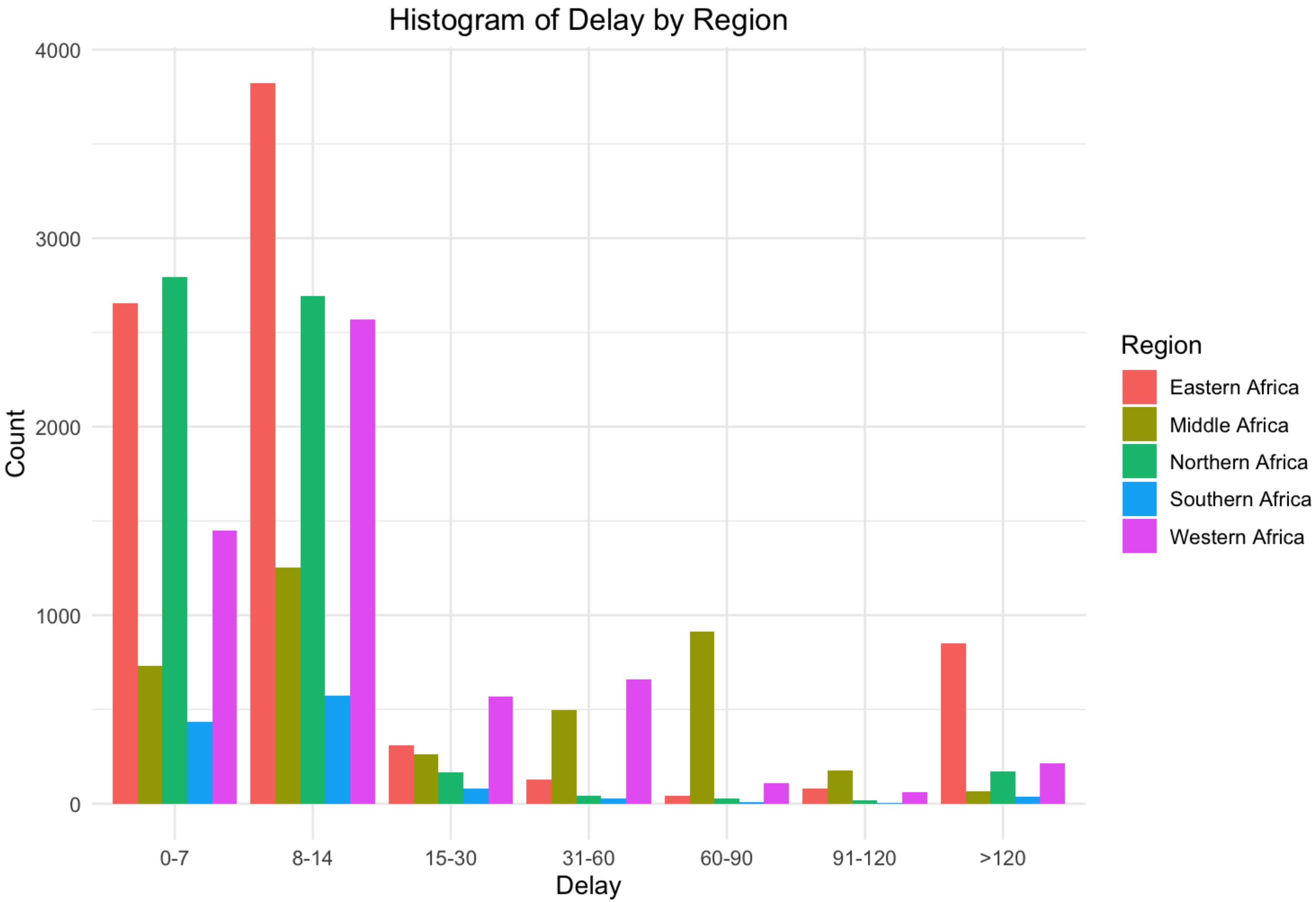
These questions will be answered better with KM but some histo won't hurt

GLOBAL VIEW : DESCRIPTIVE



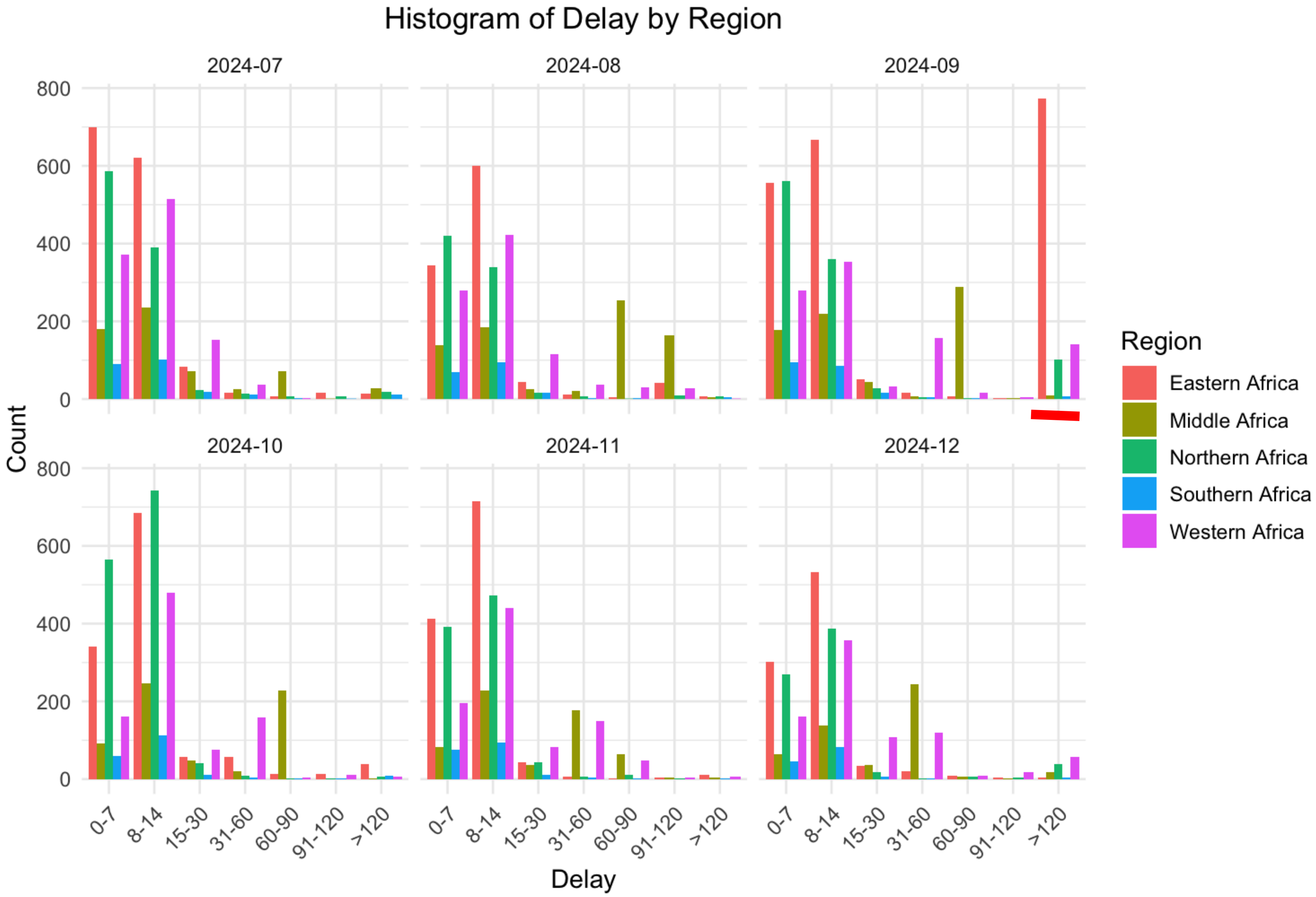
- The effect of **event type** on delay does not vary much over time
- Can we assume a constant effect? **We'll find out more with the KM**
- **What happened in sept 2024. Very different from the other months in the >120**

GLOBAL VIEW : DESCRIPTIVE



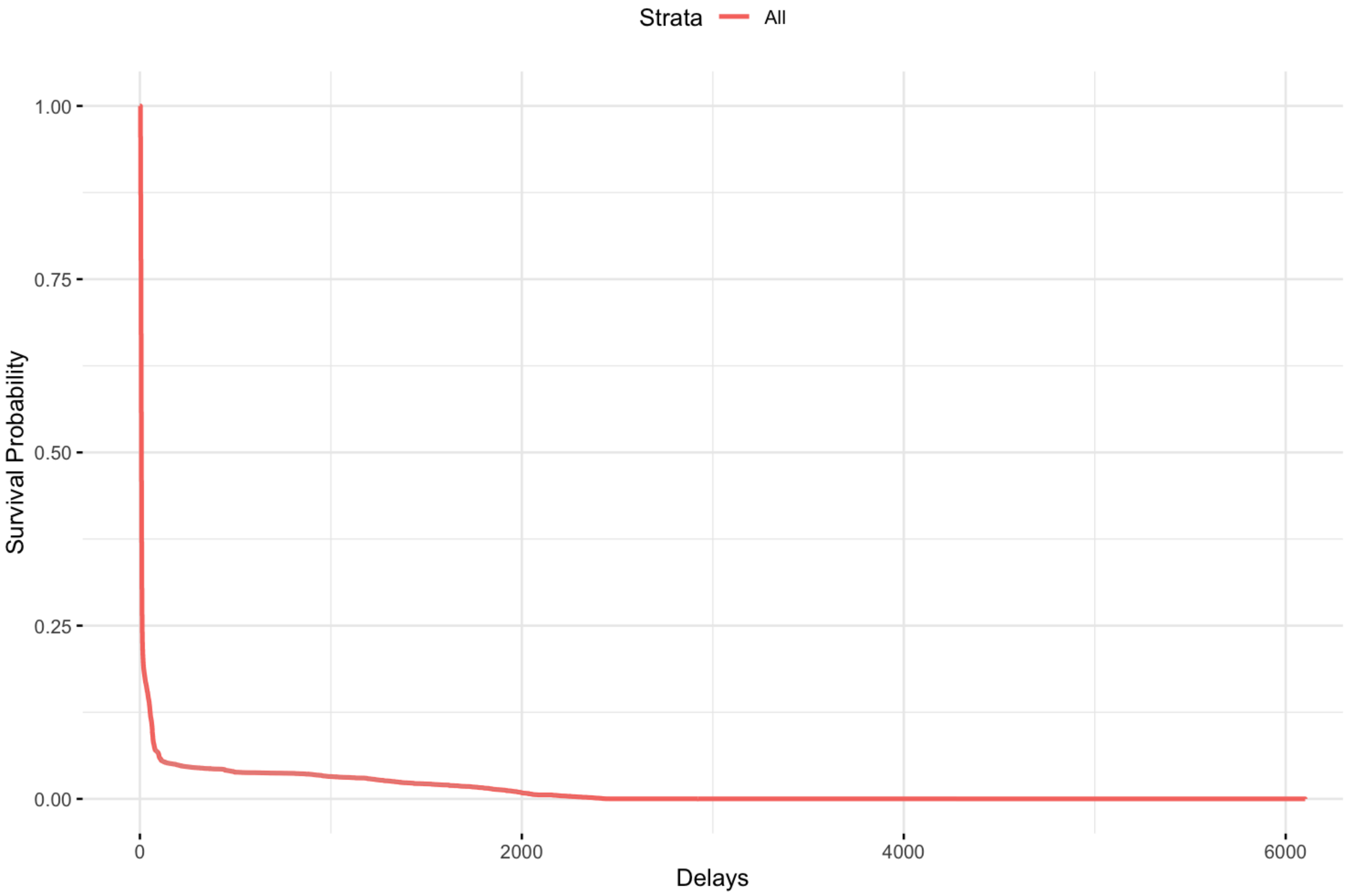
- How delays are distributed between regions?
- **Regions where there are many conflicts get report more on time?**

GLOBAL VIEW : DESCRIPTIVE



- The effect of **region** on delay does not vary much over time
- Can we assume a constant effect? **We'll find out more with the KM**

KM

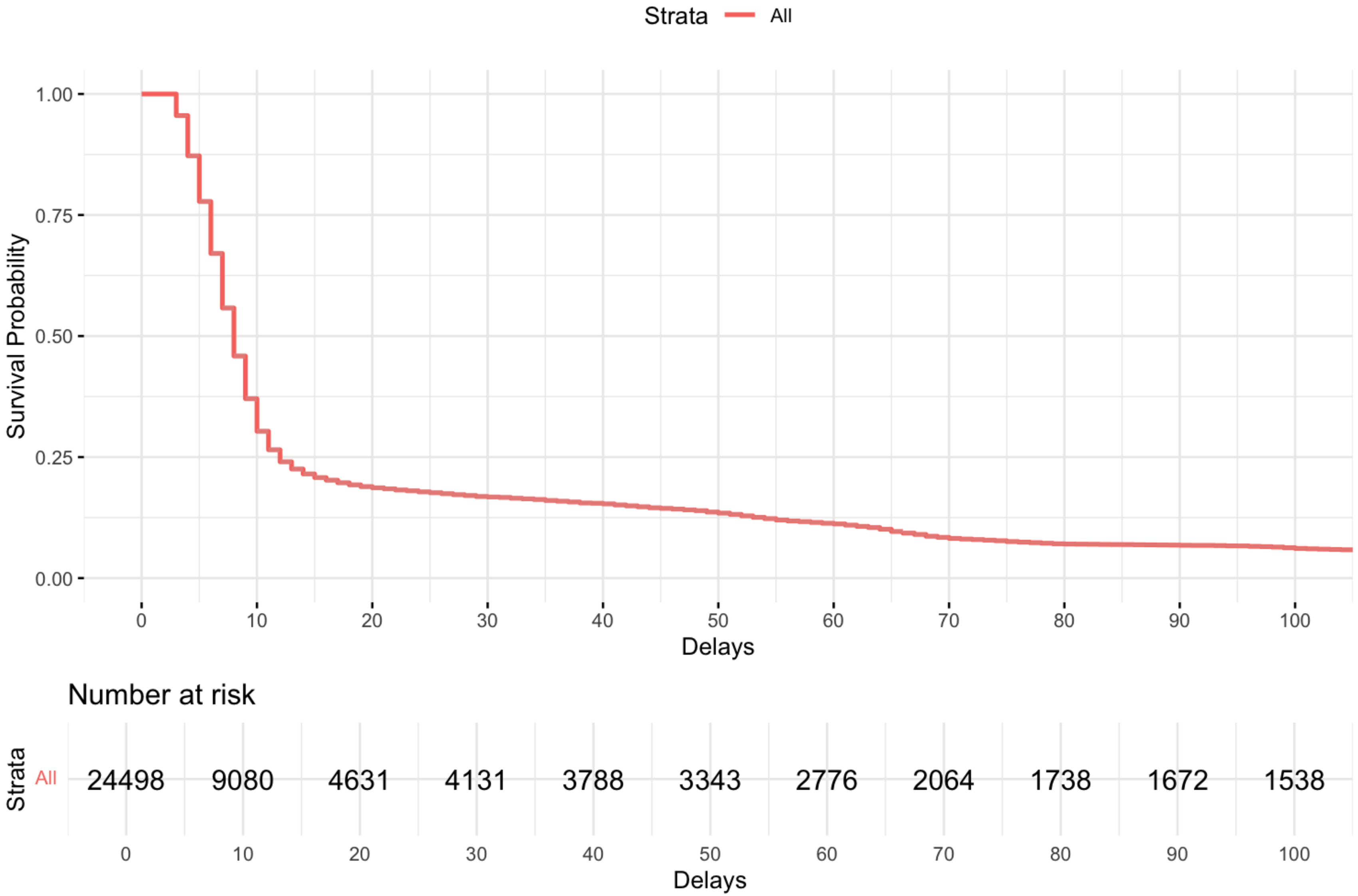


Long tail but where do we right truncate it?

The survival function decreases considerably between 0 and 80.

Short term delays are relatively more predominant

KM



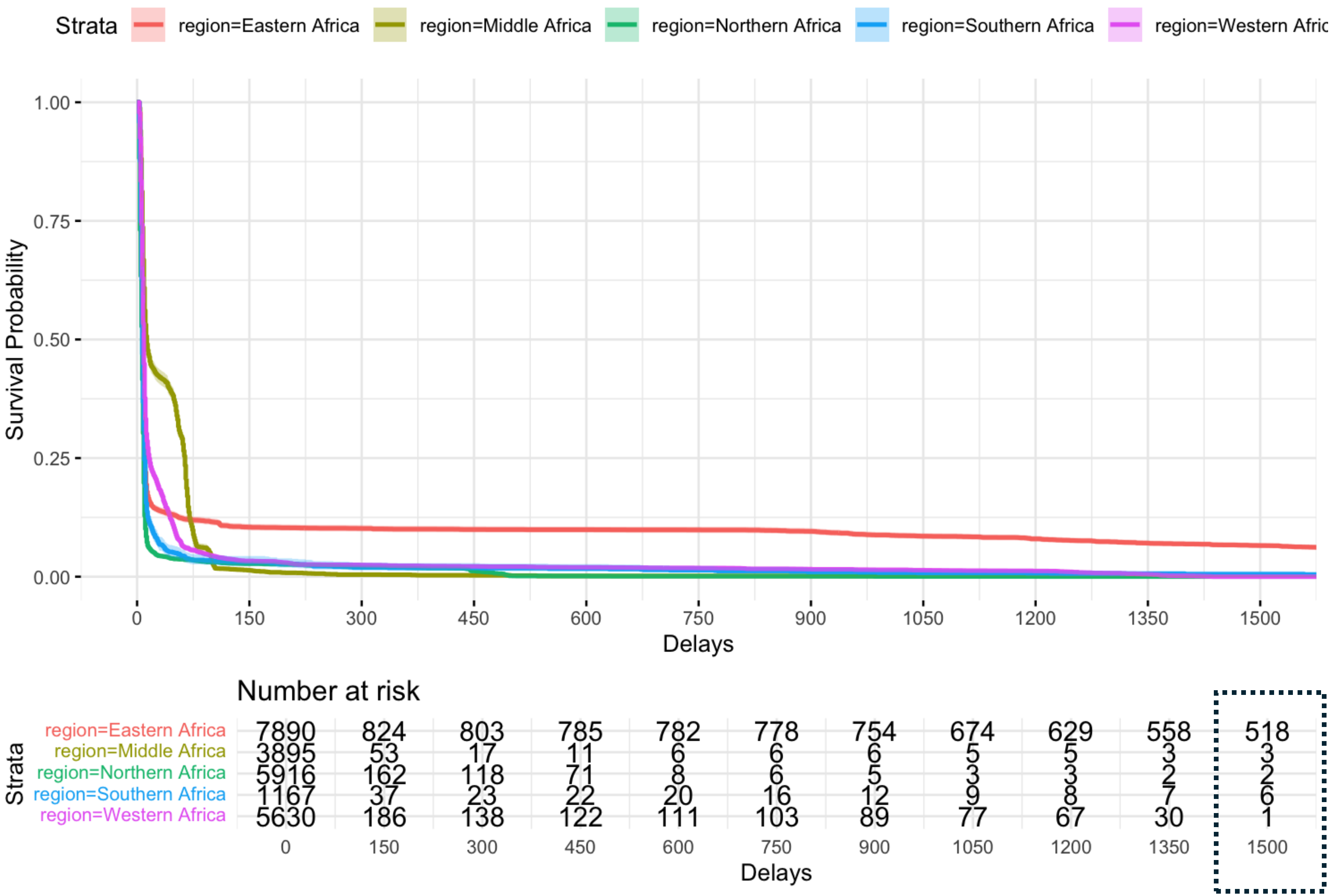
Number at risk

Strata	0	10	20	30	40	50	60	70	80	90	100
All	24498	9080	4631	4131	3788	3343	2776	2064	1738	1672	1538

S_hat(100) = 0.063

This should have been smaller if not for Somalia (East Africa)

KM (Region)



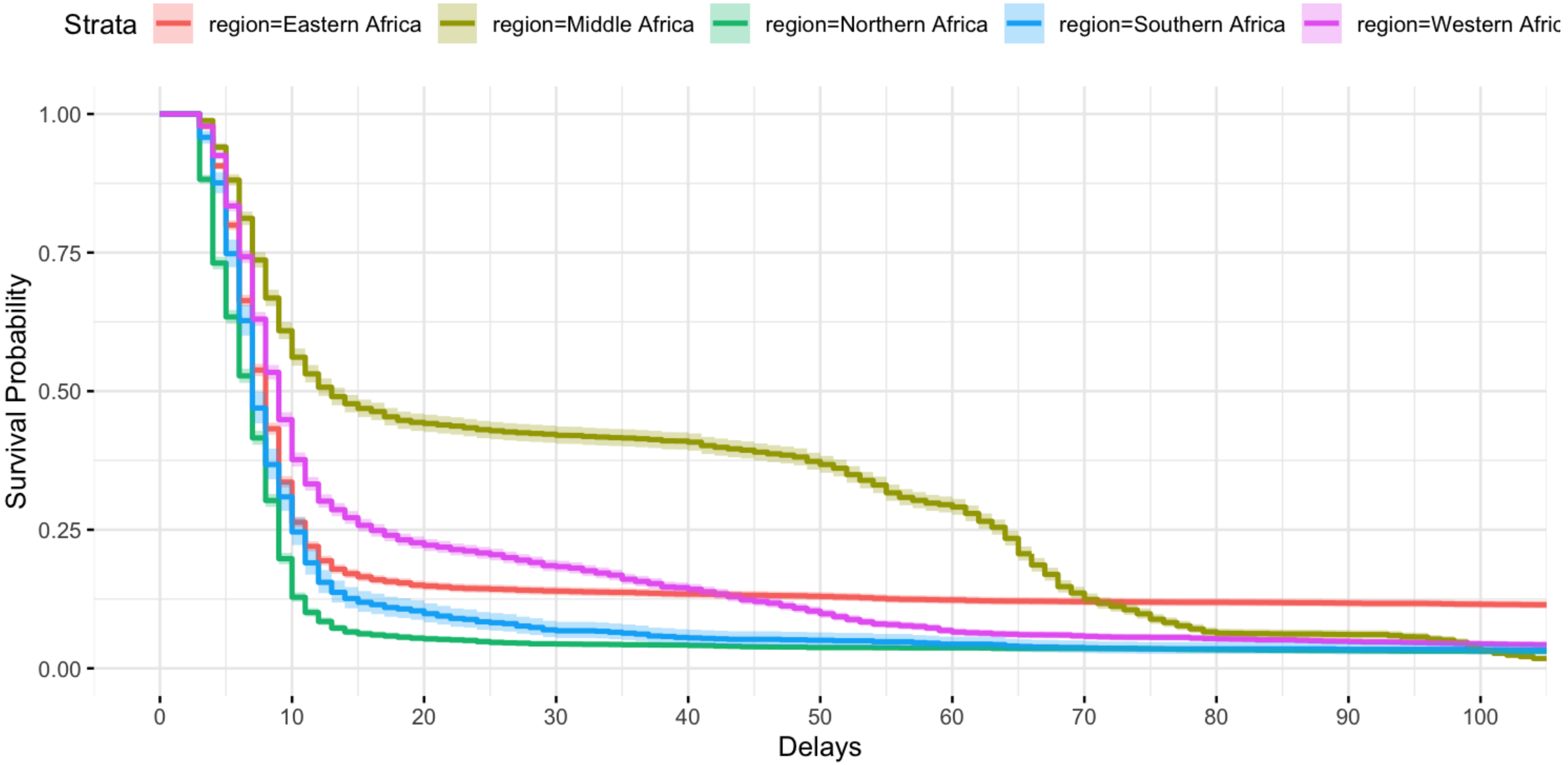
Strata	0	150	300	450	600	750	900	1050	1200	1350	1500
region=Eastern Africa	7890	824	803	785	782	778	754	674	629	558	518
region=Middle Africa	3895	53	17	11	6	6	6	5	5	3	3
region=Northern Africa	5916	162	118	71	8	6	5	3	3	2	2
region=Southern Africa	1167	37	23	22	20	16	12	9	8	7	6
region=Western Africa	5630	186	138	122	111	103	89	77	67	30	1

- East Africa goes down very fast between 0 and 14, and then a steady long tail.

- How fast do these curves descend?

- **The effect of the REGION on the Hazard function is less for higher delay?**

KM (Region)



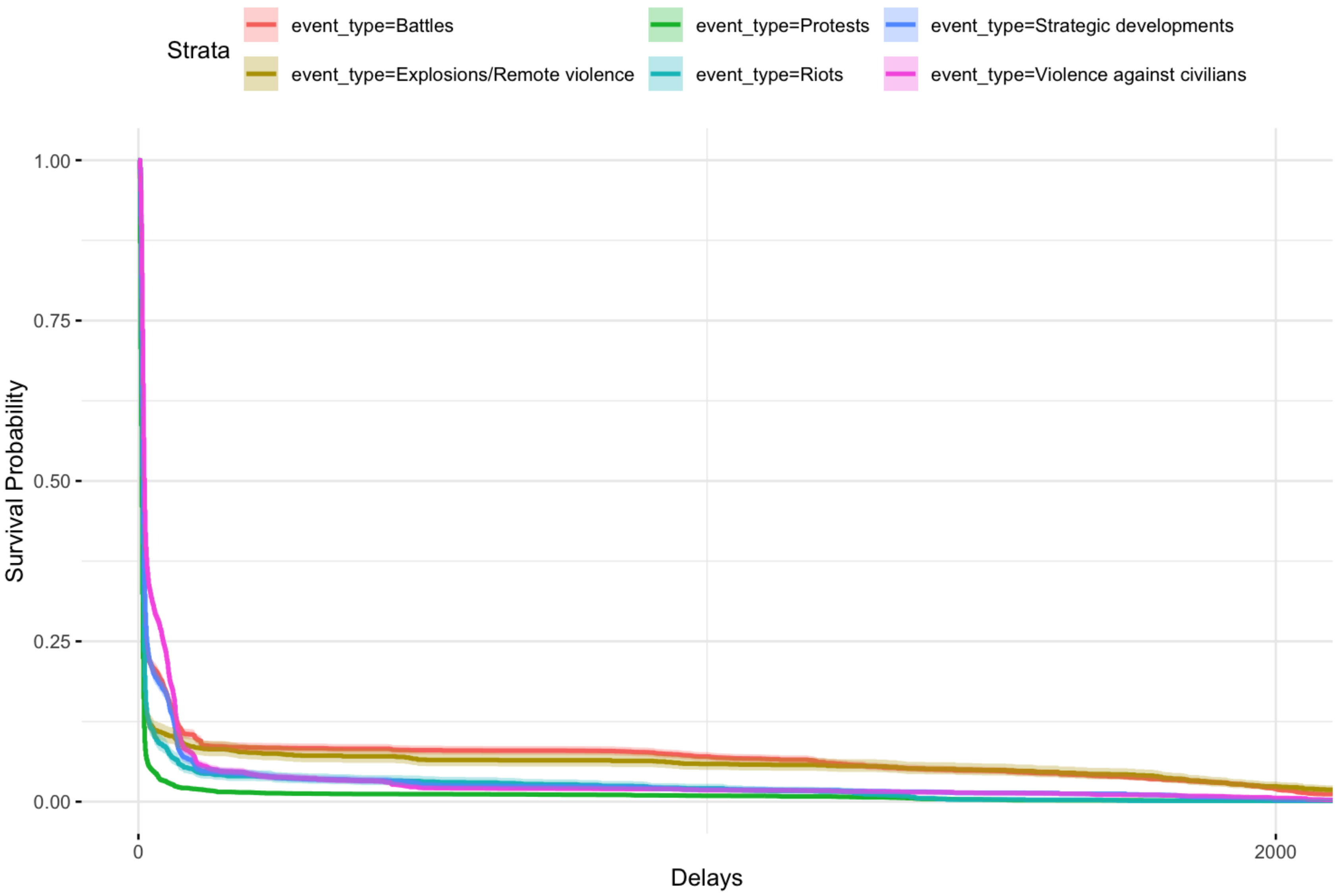
Number at risk

Strata	0	10	20	30	40	50	60	70	80	90	100
region=Eastern Africa	7890	2651	1184	1101	1058	1026	973	949	940	929	913
region=Middle Africa	3895	2372	1728	1645	1597	1454	1150	529	256	238	152
region=Northern Africa	5916	1170	322	261	248	222	218	212	199	190	184
region=Southern Africa	1167	361	121	81	65	59	51	44	41	40	39
region=Western Africa	5630	2526	1276	1043	820	582	384	330	302	275	250

Delays

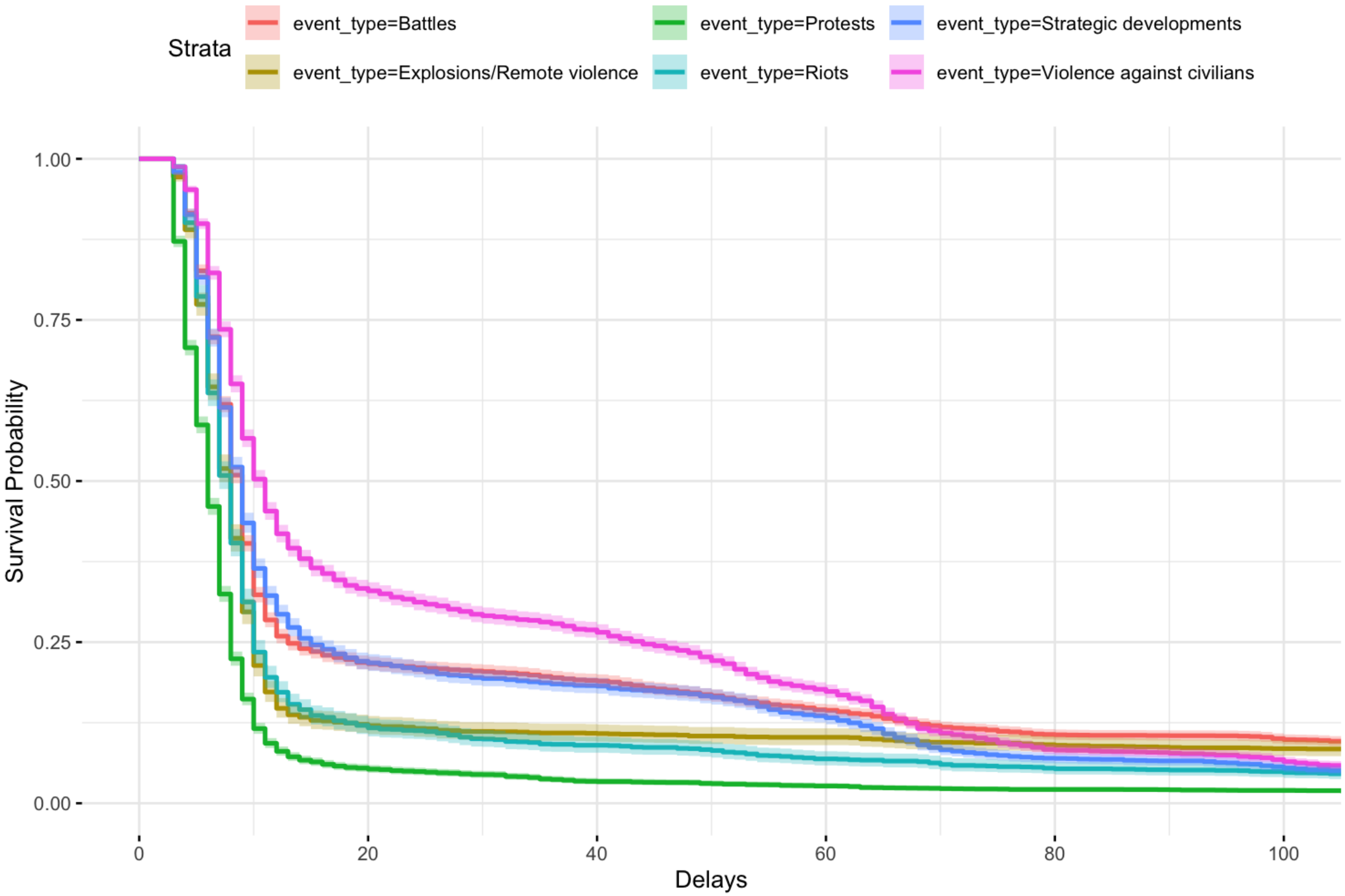
- Northern Africa has high number at risk but still decreases very fast
- Middle Africa has the slowest decrease while having the second least number at risk
- East Africa has the highest number at risk and decreases fast before staying steady (because of Somalia)
- **Is there an exponential decay of the number at risk? (Decrease that is function of the amount)**

KM (Events)



- We observe significant difference between the group (Battles, Explosion) and the rest. We'll also see later that there is also a higher error of event classification in the first group because they are very related and hence more misclassified [battles that are misclassified as explosion and almost all of them in Somalia]

KM (Events)



Strata
event_type=Battles
event_type=Explosions/Remote violence
event_type=Protests
event_type=Riots
event_type=Strategic developments
event_type=Violence against civilians
Survival Probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
100
Delays

What to discuss?

- **Is early truncation then a very smart move here?**
- Unlike the focus on short-term analysis seen in pandemic research, conflict studies require equal attention to both short-term and long-term analyses. This dual focus might be imp for capturing the complex dynamics of conflict over varying timeframes (CYCLE OF CONFLICT vs CYCLE OF PANDEMIC)
- All of the previous results will analytically make more sense when we model the Hazard Function

Measurement Error effect

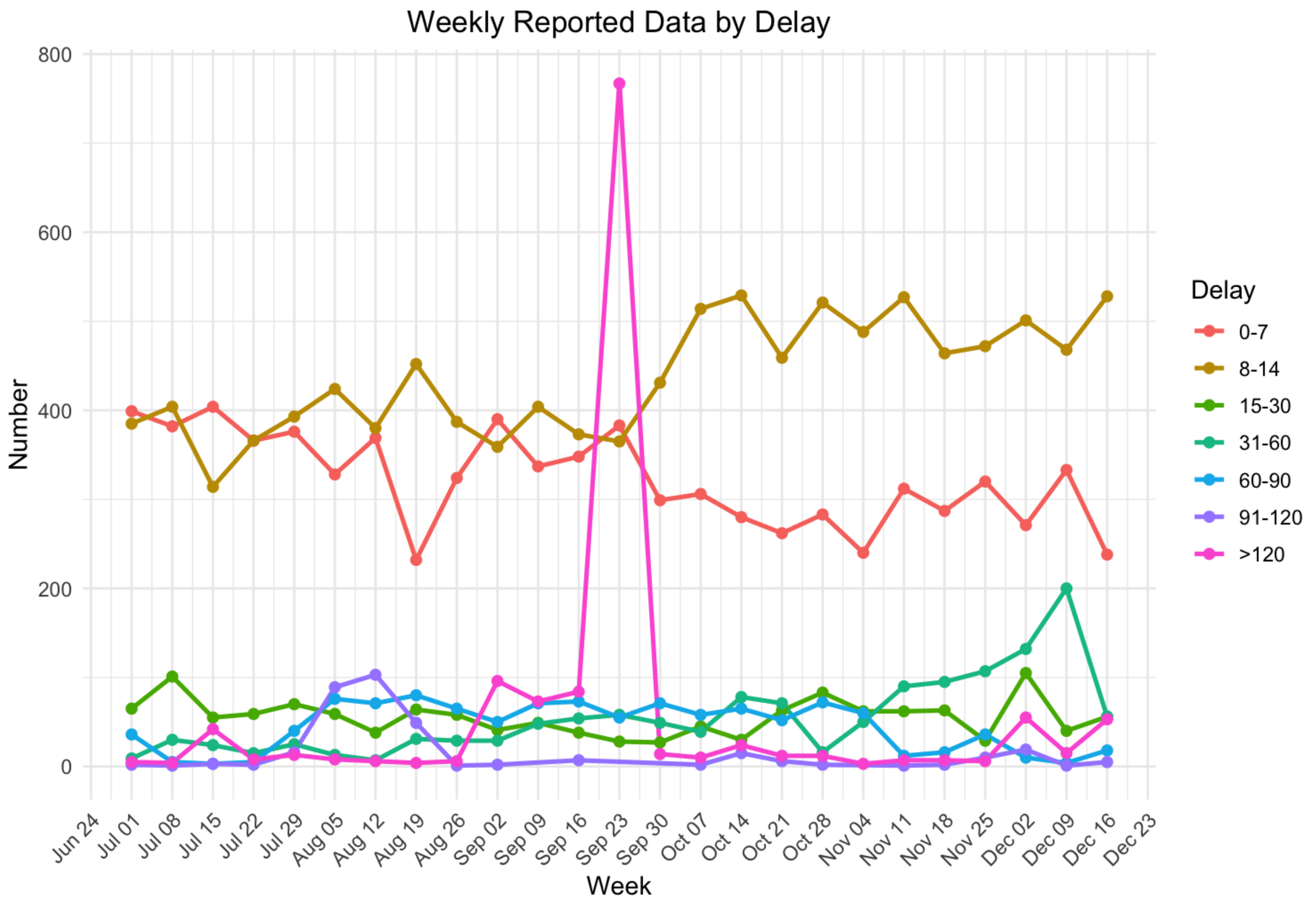
- **The measurement error might introduce bias when we estimate the effect of the covariates on delay**

- **How does it change the survival function?**

- **How will we also formalize it into our model? It changes the time at which the true event is reported (possibility: add more events in the class of events with longer delay and removes events from the class of events with shorter delay)**

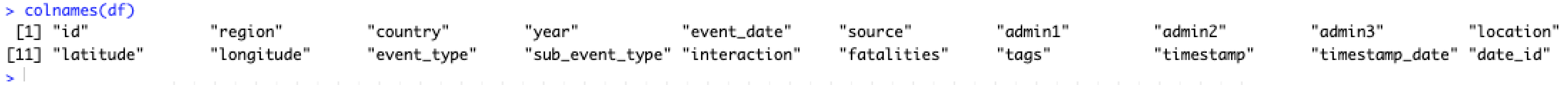
Error effect : how do we define it? (TBR)

NO ERROR ASSUMP

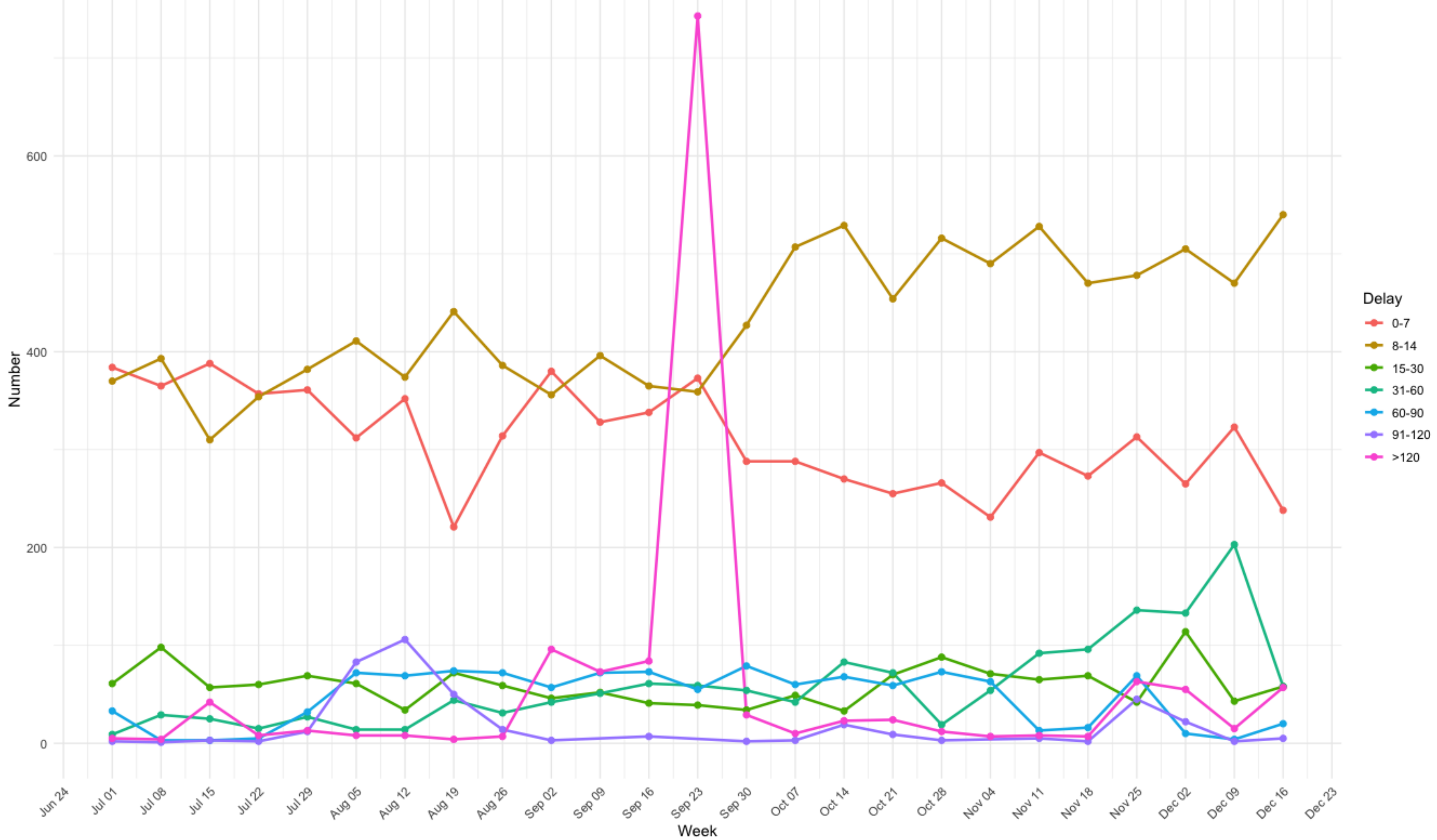


Error effect on delays

ANY CHANGE IN ONE OF THE COLUMNS EXCEPT SOURCE

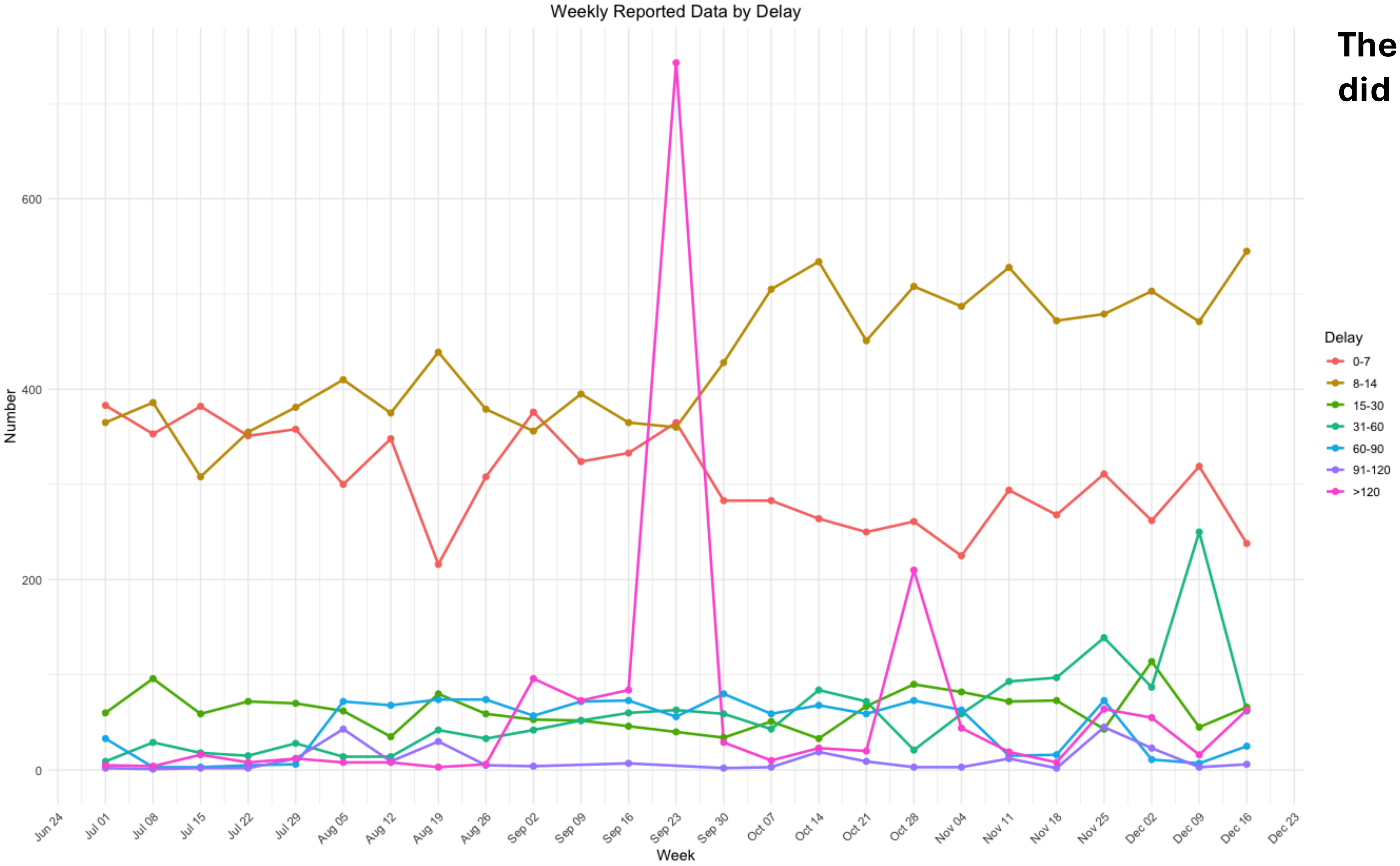


```
> colnames(df)
 [1] "id"          "region"      "country"     "year"          "event_date"  "source"      "admin1"      "admin2"      "admin3"         "location"
[11] "latitude"    "longitude"   "event_type"  "sub_event_type" "interaction" "fatalities"  "tags"        "timestamp"   "timestamp_date" "date_id"
> 
```



Error effect on delays

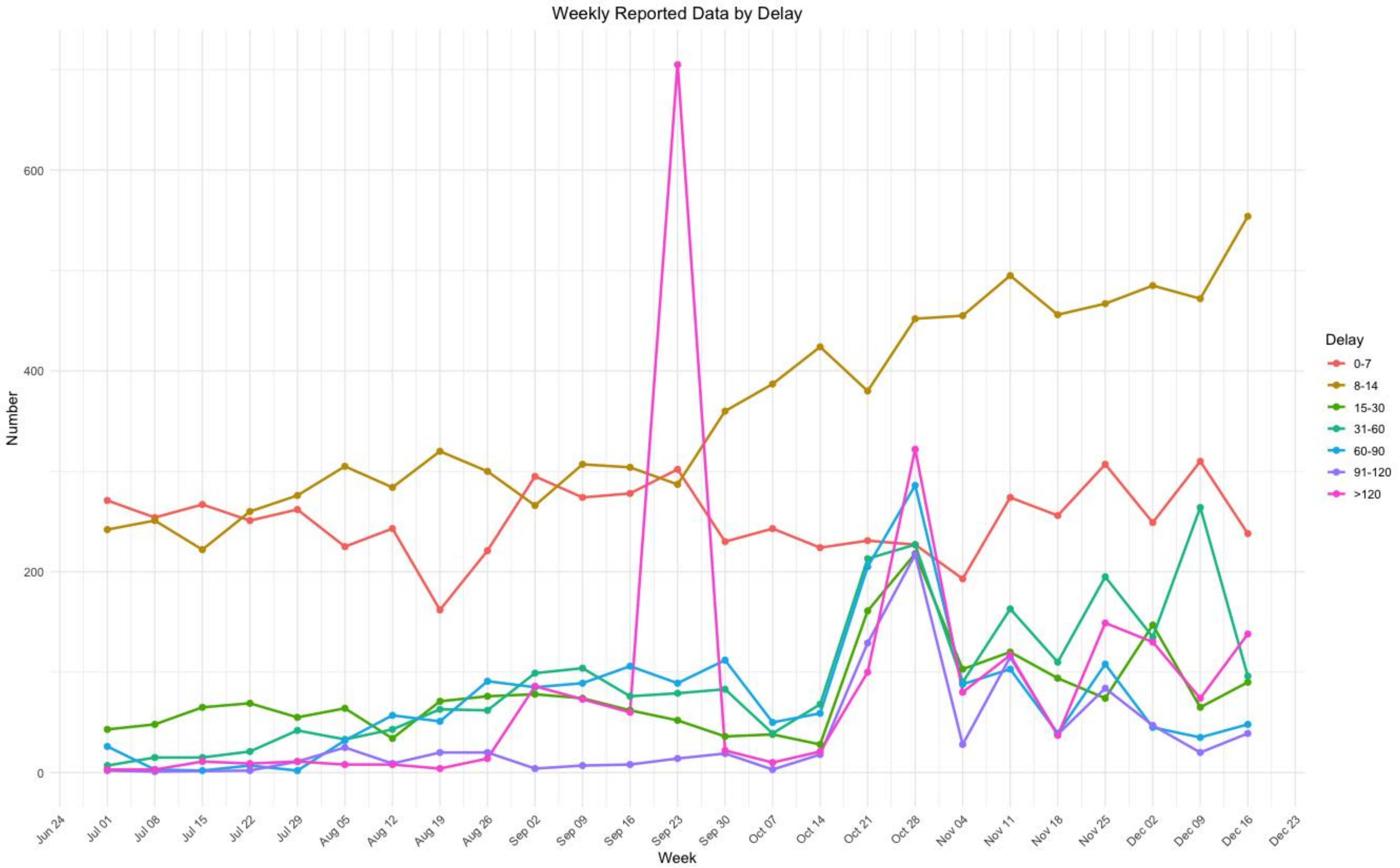
ANY CHANGE IN ONE OF THE COLUMNS



The events with short delay did not change

Error effect on delays

THE LATEST TIMESTAMP

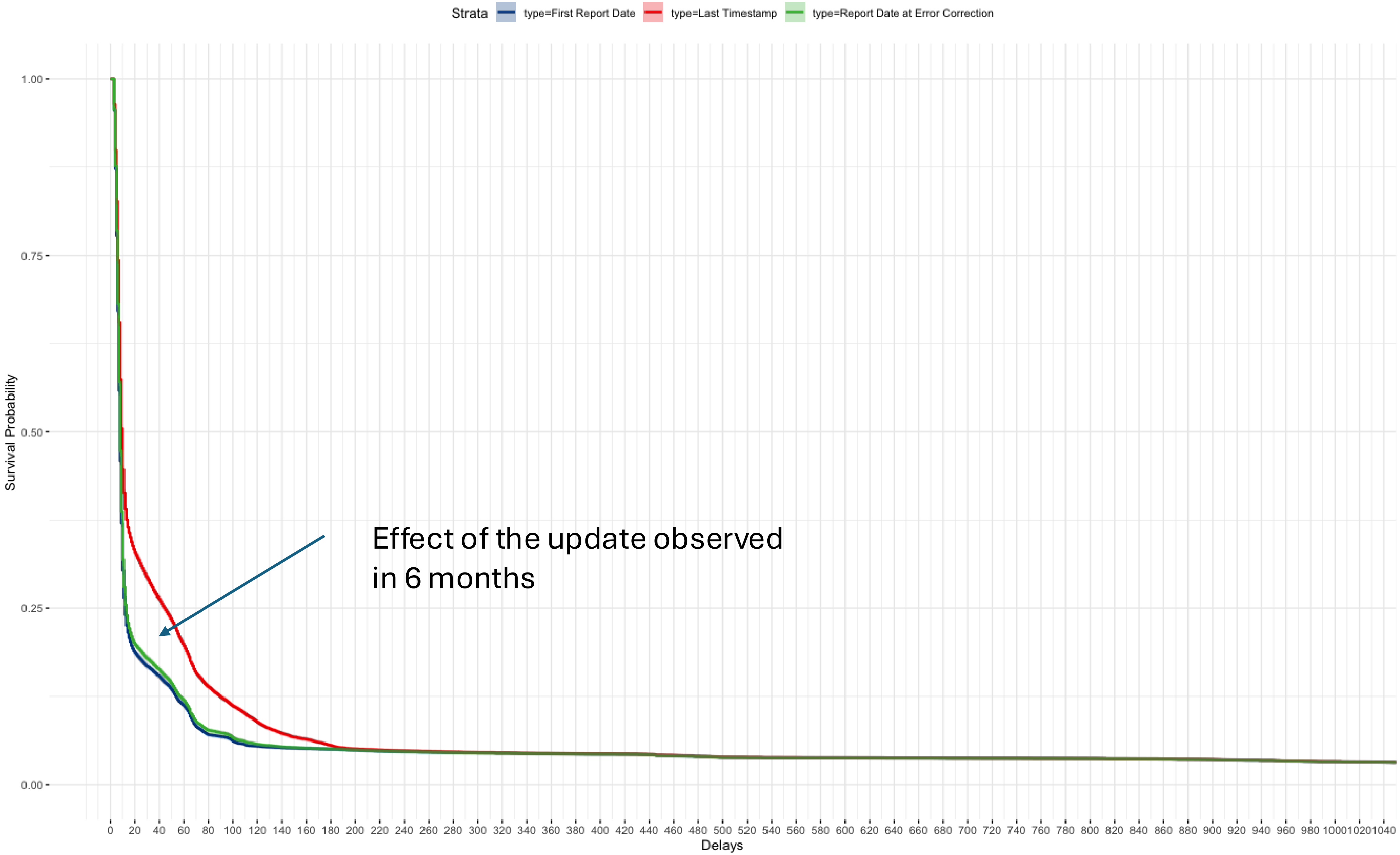


Error effect on delays

- How our KM will change depend on how we define the error (what kind of update is considered as error correction : change of event for example, change of location that is more than 60km,…??)

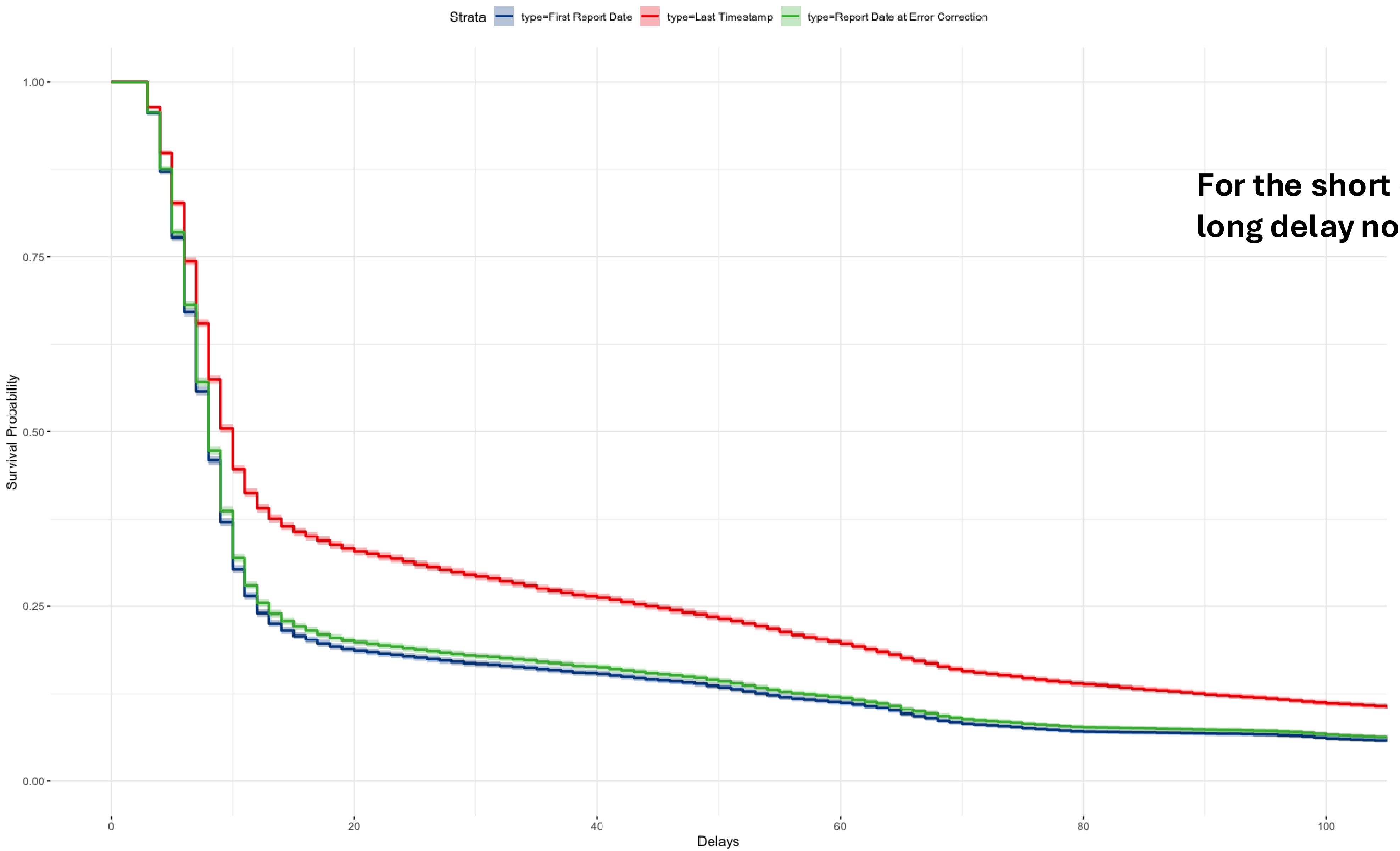
- The columns not figured in our data are not relevant?

Error effect on delays



Effect of the update observed in 6 months

Error effect on delays



For the short delay and long delay not very diff

A bit of discussion

- Are these differences significant? Is it worth modelling? (Error observed on data collected in 6 months, is there a way to foresee how error grows with higher period of observation)

- Will we tackle this hierarchically? (nowcast model then error correction based on it, or error correction then nowcast model based on the corrected data)

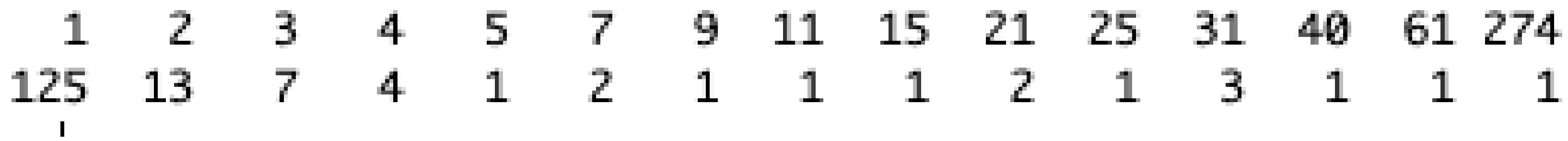
- We observe very few error corrections among newly reported events because we observed the change for only 6 months. One extreme case that we would fail to notice with such data is a negative relationship between delay of report and delay of correction from report (**event with long delay gets corrected rapidly and event with short delay gets corrected after a long time**)

- Can we gain knowledge about the error correction from the updates? (are they proportional)

- How long does correction take place?

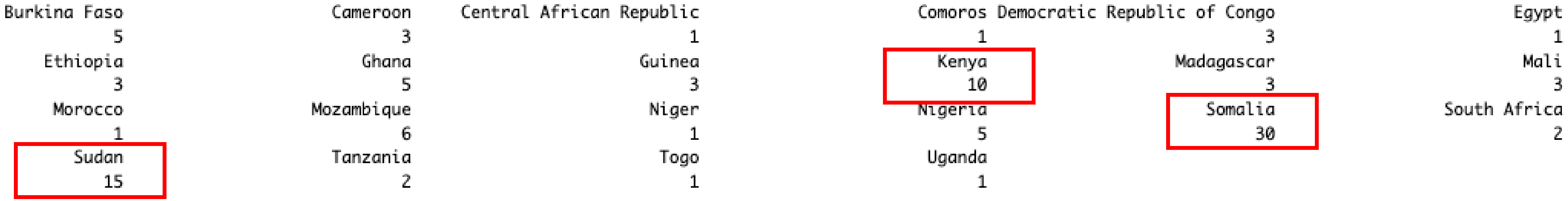
Observational Error: Descri

- 164 changes in **date event** observed between first and following downloaded datasets (up to 6 months)



```
Diff    1   2   3   4   5   7   9  11  15  21  25  31  40  61 274
N     125  13   7   4   1   2   1   1   1   2   1   3   1   1   1
```

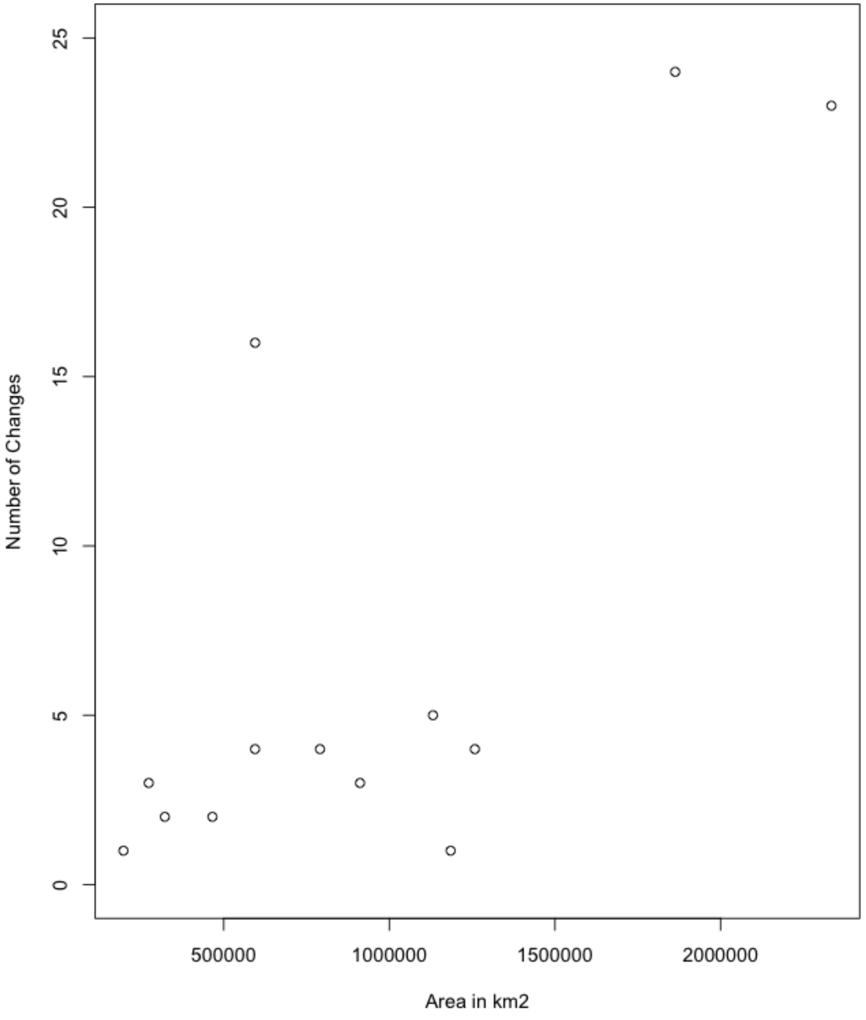
- 105 changes in **event type** observed in 6 months

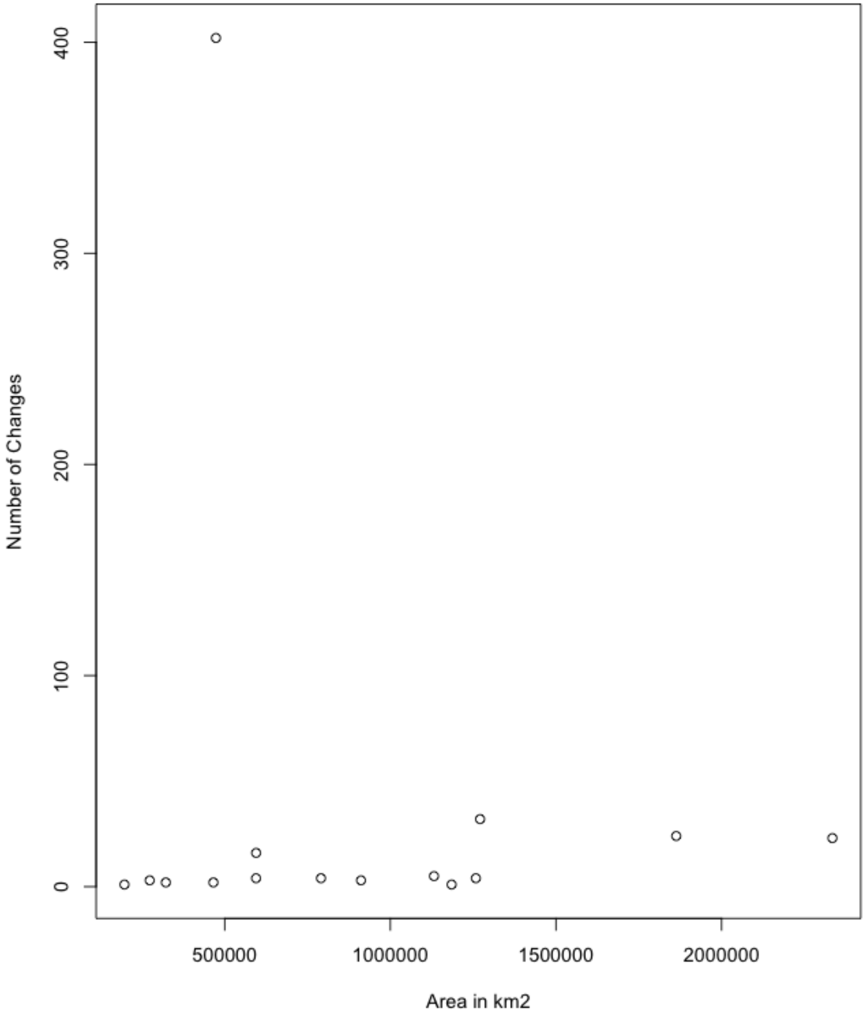


```
Burkina Faso            Cameroon   Central African Republic             Comoros Democratic Republic of Congo          Egypt
           5                   3                          1                   1                                3              1
    Ethiopia               Ghana                     Guinea               Kenya                       Madagascar           Mali
           3                   5                          3                  10                                3              3
     Morocco          Mozambique                      Niger             Nigeria                          Somalia   South Africa
           1                   6                          1                   5                               30              2
       Sudan            Tanzania                       Togo              Uganda
          15                   2                          1                   1
```

Interesting fact : The level of corruption in countries with relatively high event changes is also relatively higher

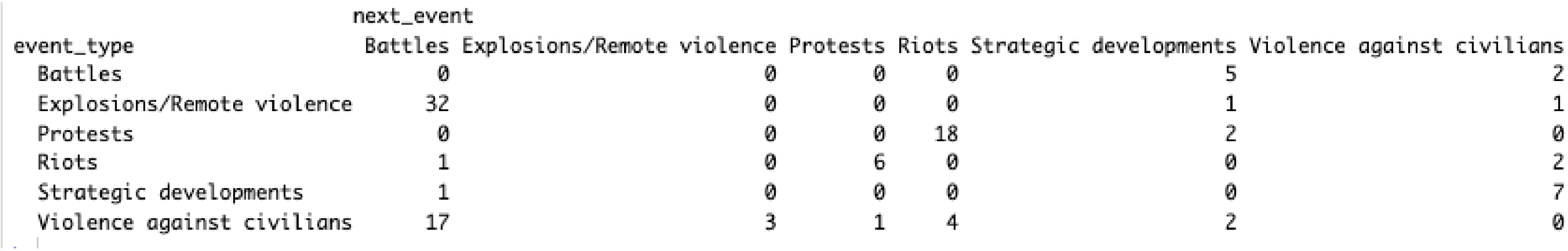
Observational Error: Descri





Change of location (more than a radius of 60km)
Somalia

Observational Error: Descri

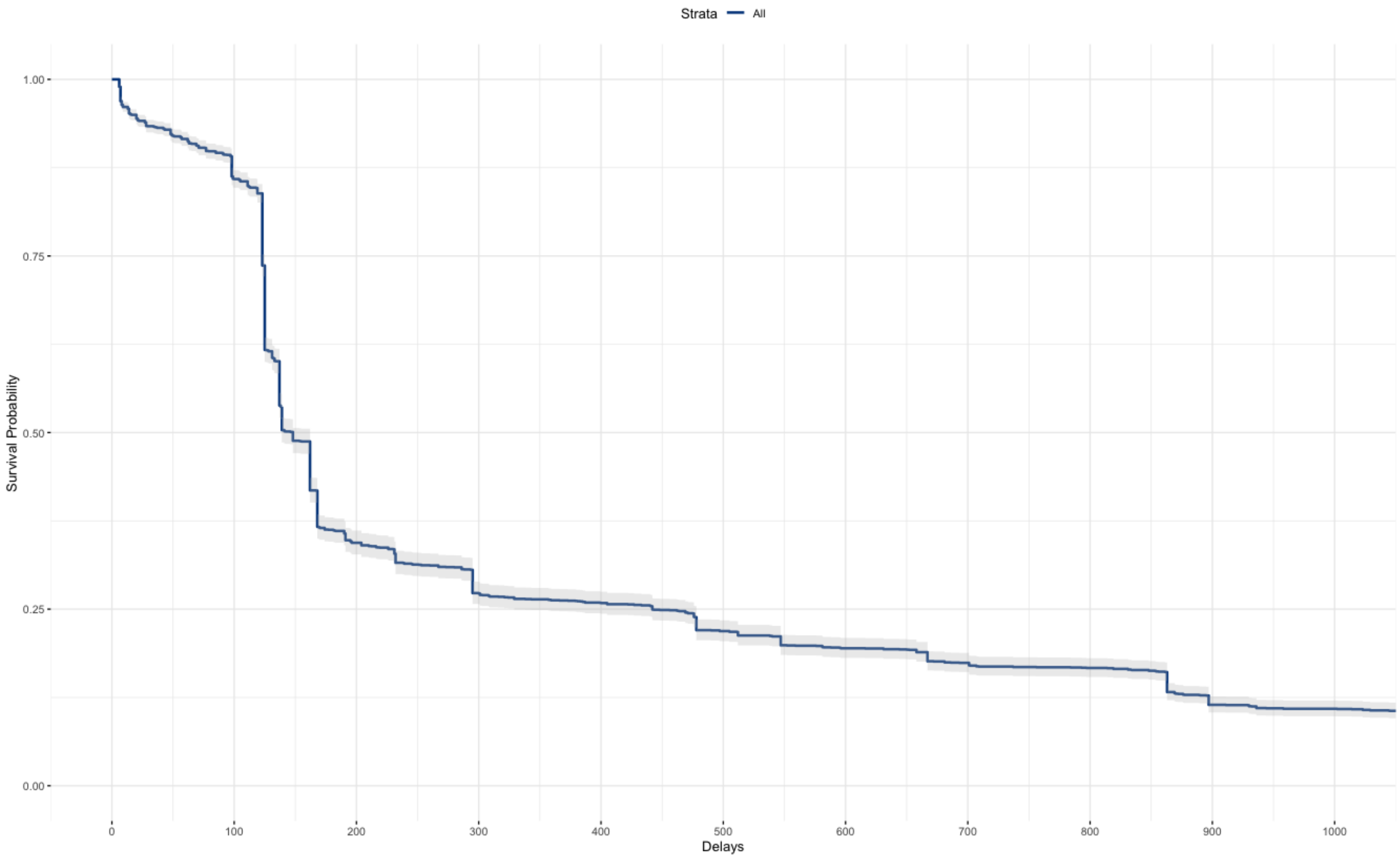


```
                            next_event
event_type                   Battles Explosions/Remote violence Protests Riots Strategic developments Violence against civilians
  Battles                          0                          0        0     0                      5                          2
  Explosions/Remote violence      32                          0        0     0                      1                          1
  Protests                         0                          0        0    18                      2                          0
  Riots                            1                          0        6     0                      0                          2
  Strategic developments           1                          0        0     0                      0                          7
  Violence against civilians      17                          3        1     4                      2                          0
```

28 out of the 32 in Somalia

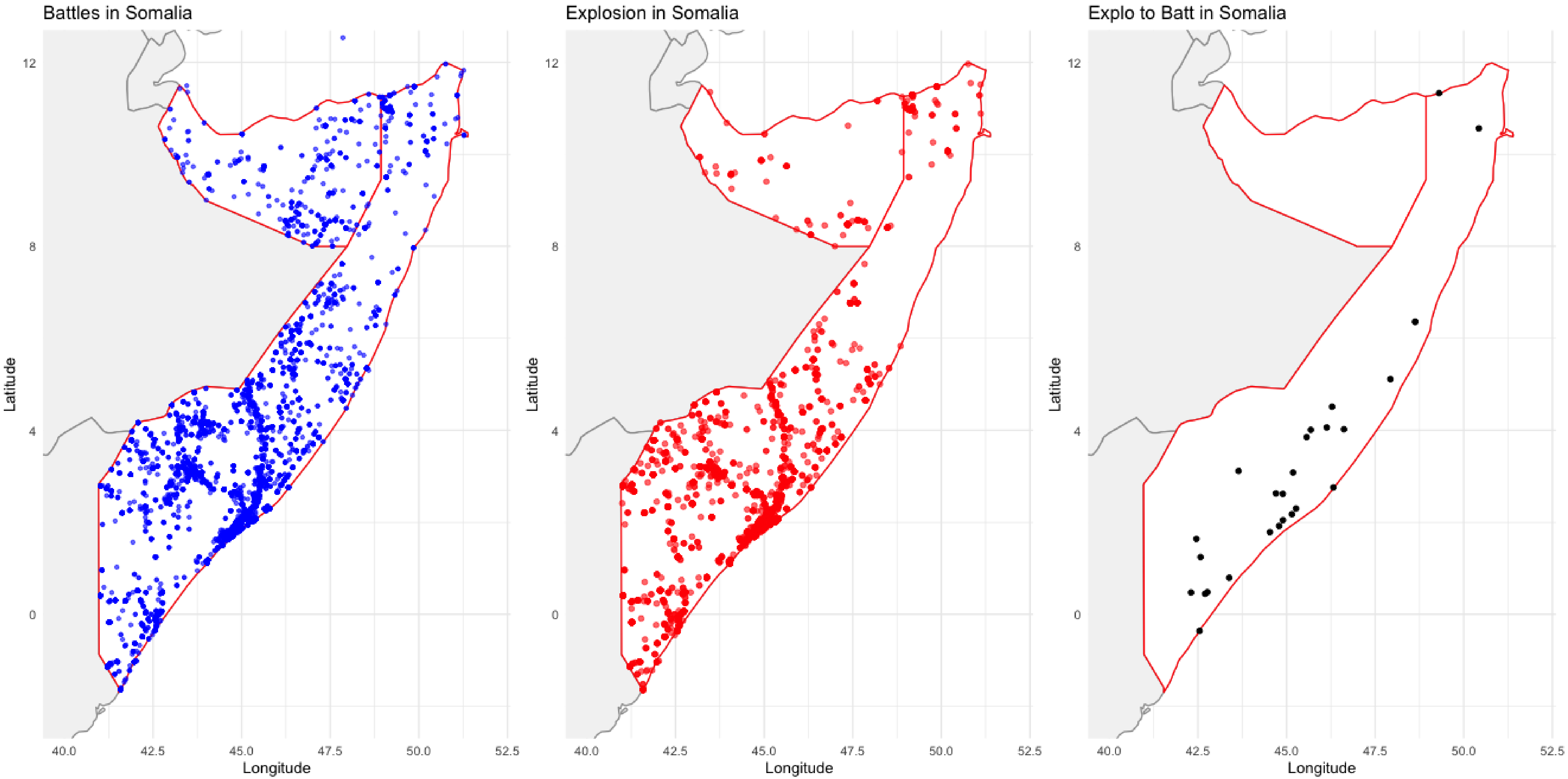
These corrections make sense. Events with higher correction happen to be more correlated

How long do error get corrected



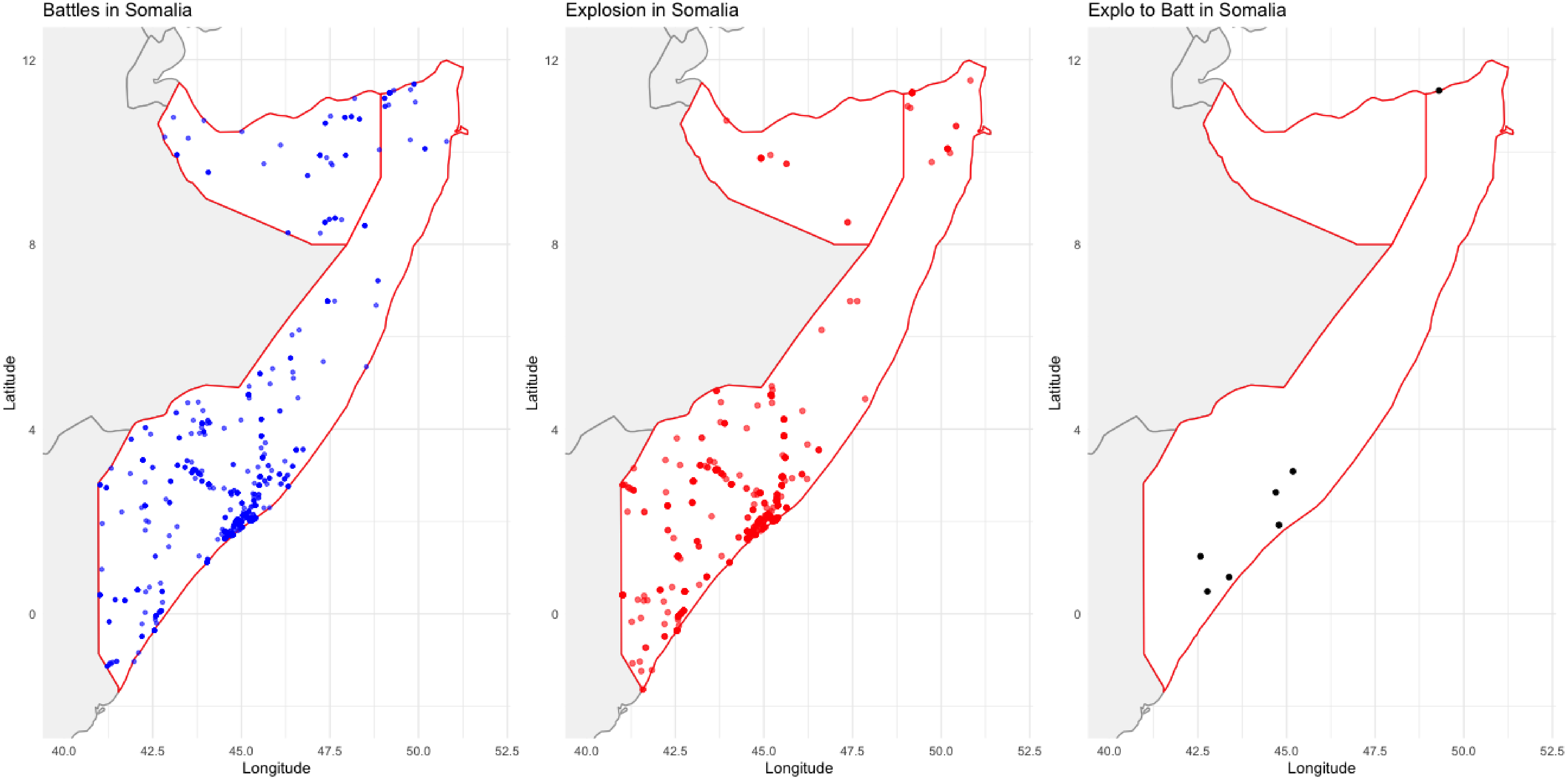
Is there a pattern we can pick in the error correction (something our model can statistically learn)

IN GENERAL



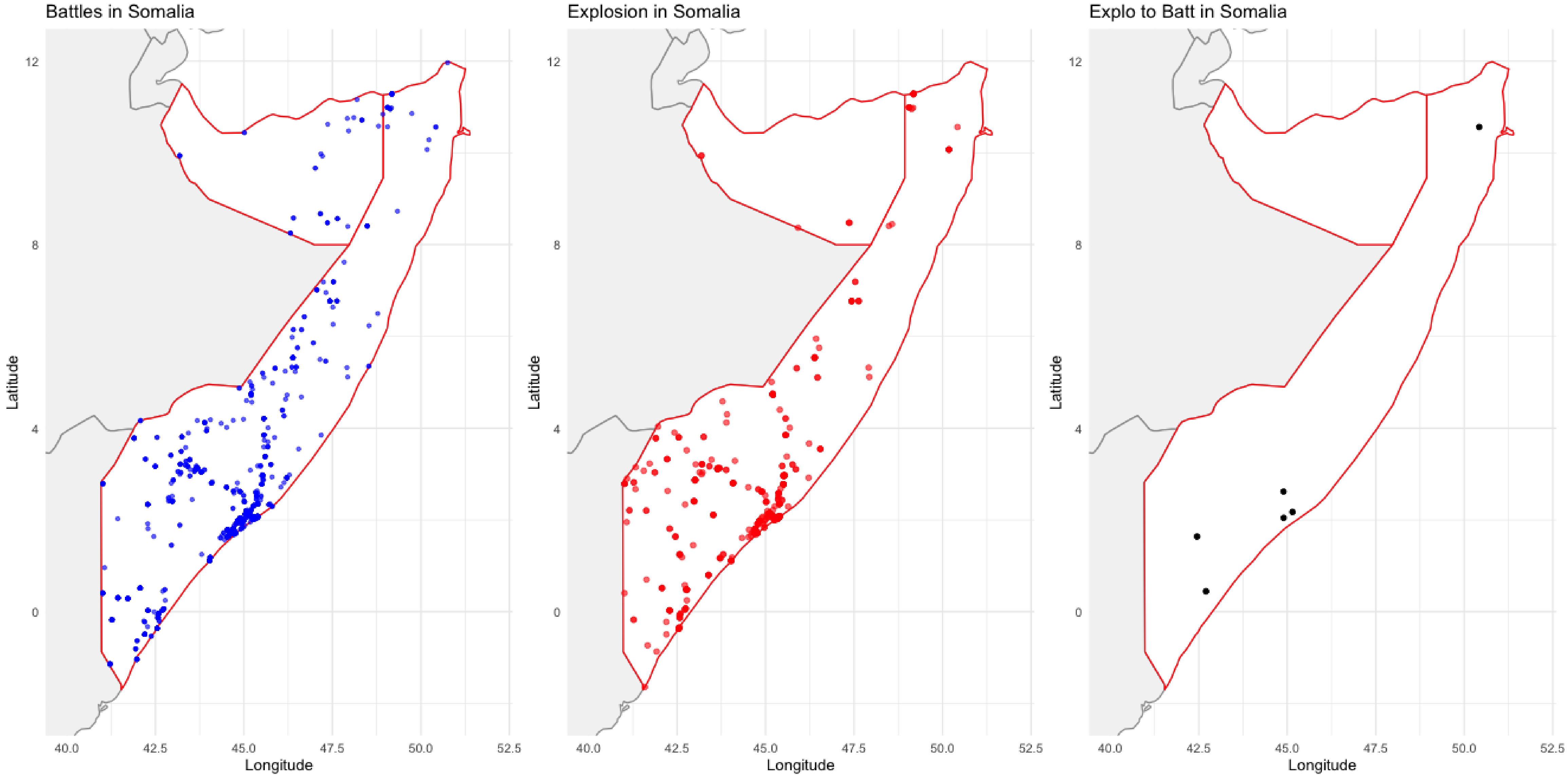
Battles in Somalia
Explosion in Somalia
Explo to Batt in Somalia
Latitude
12
8
4
0
40.0
42.5
45.0
47.5
50.0
52.5
Longitude

2019



Battles in Somalia
Explosion in Somalia
Explo to Batt in Somalia
12
8
4
0
Latitude
40.0
42.5
45.0
47.5
50.0
52.5
Longitude

2020

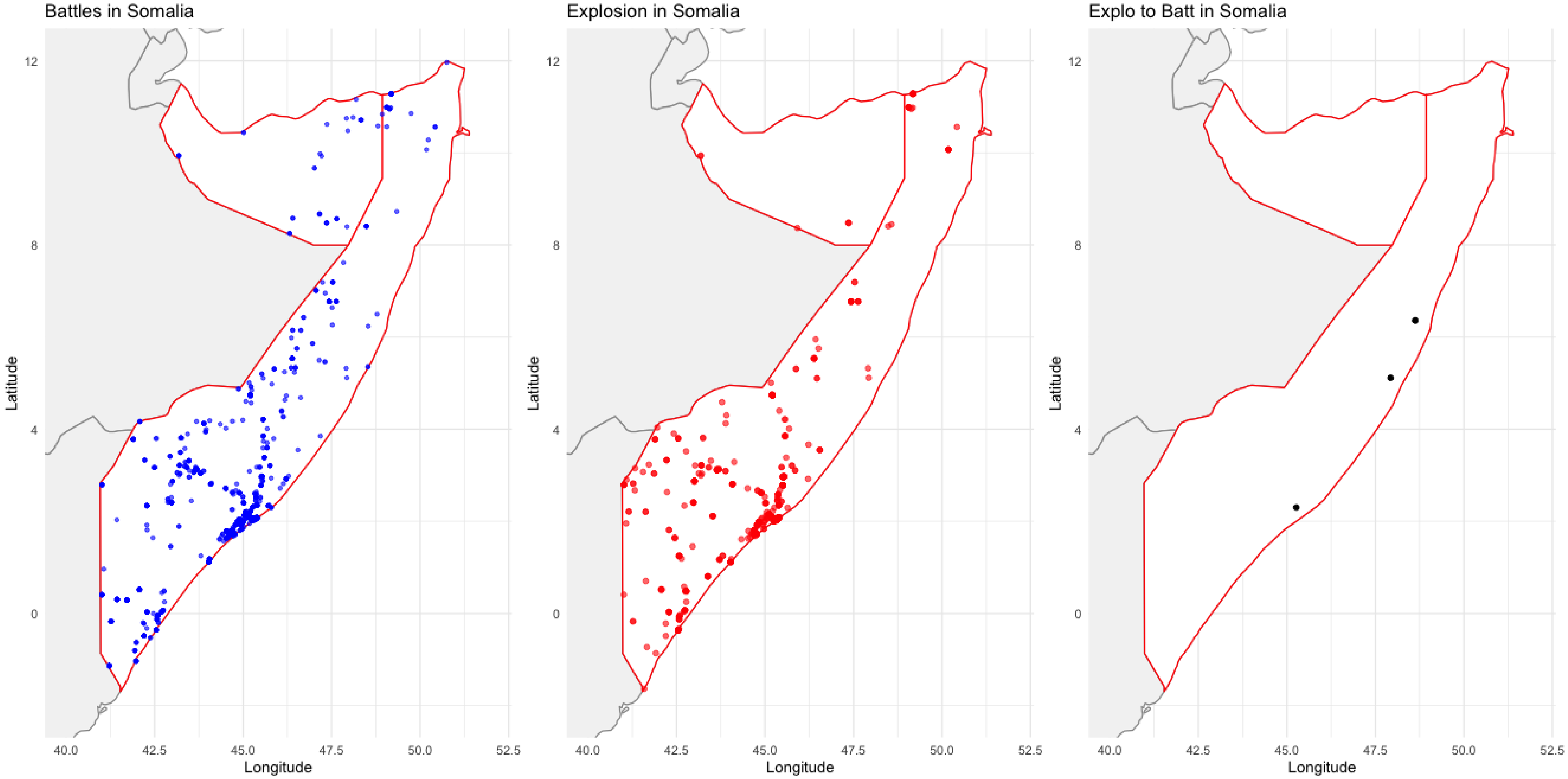


Battles in Somalia

Explosion in Somalia

Explo to Batt in Somalia

2021

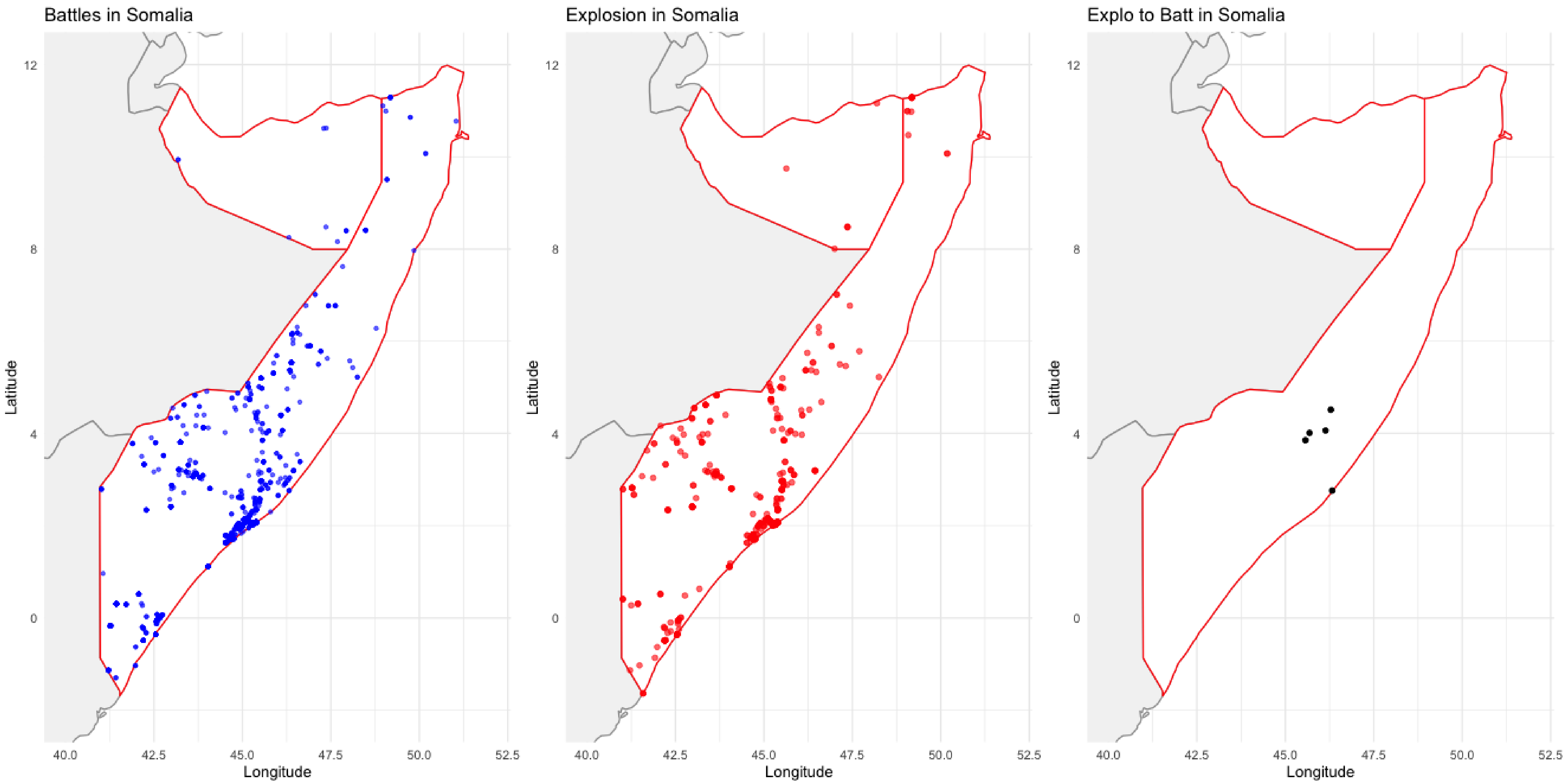


Battles in Somalia

Explosion in Somalia

Explo to Batt in Somalia

2022

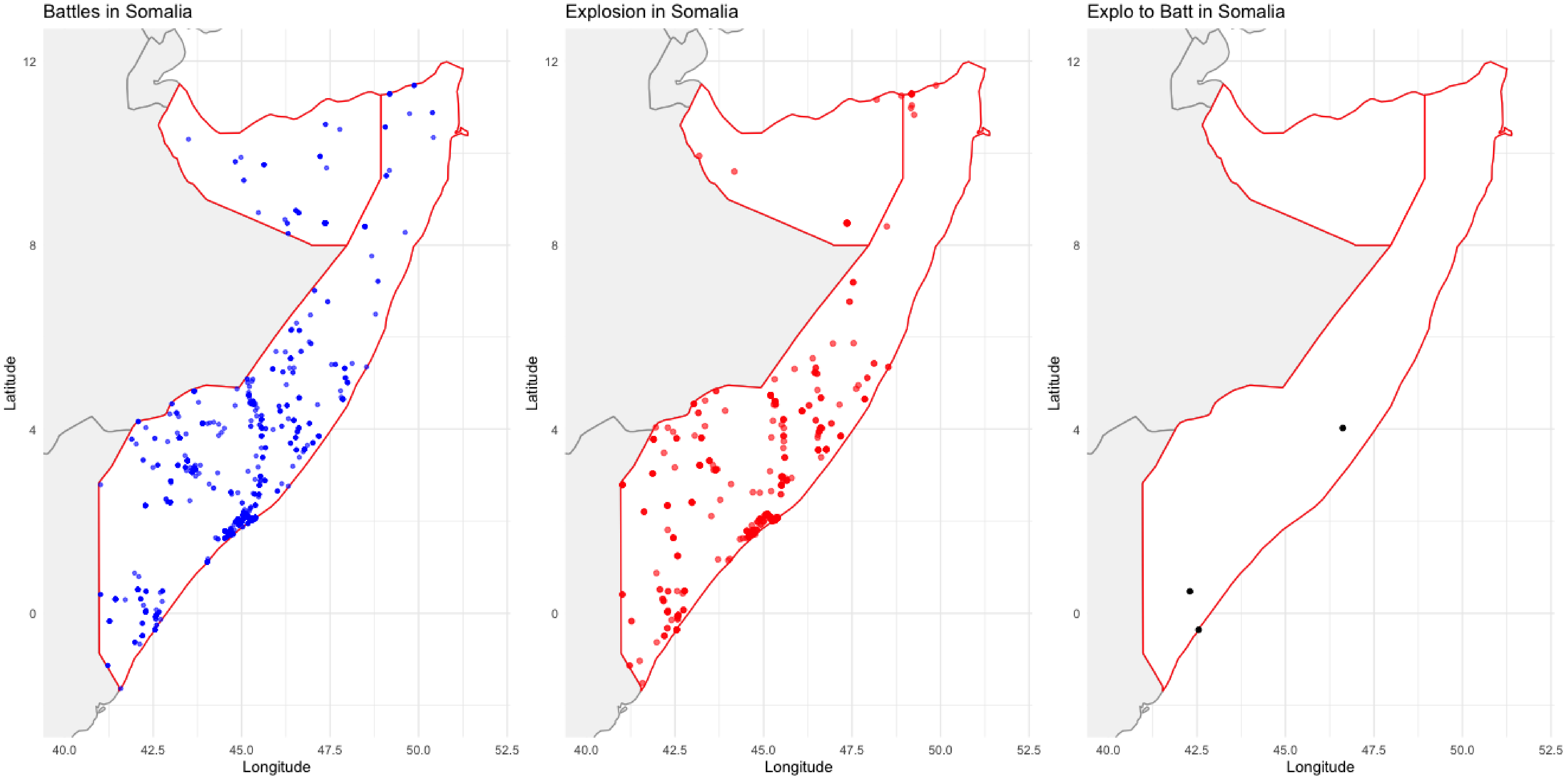


Battles in Somalia

Explosion in Somalia

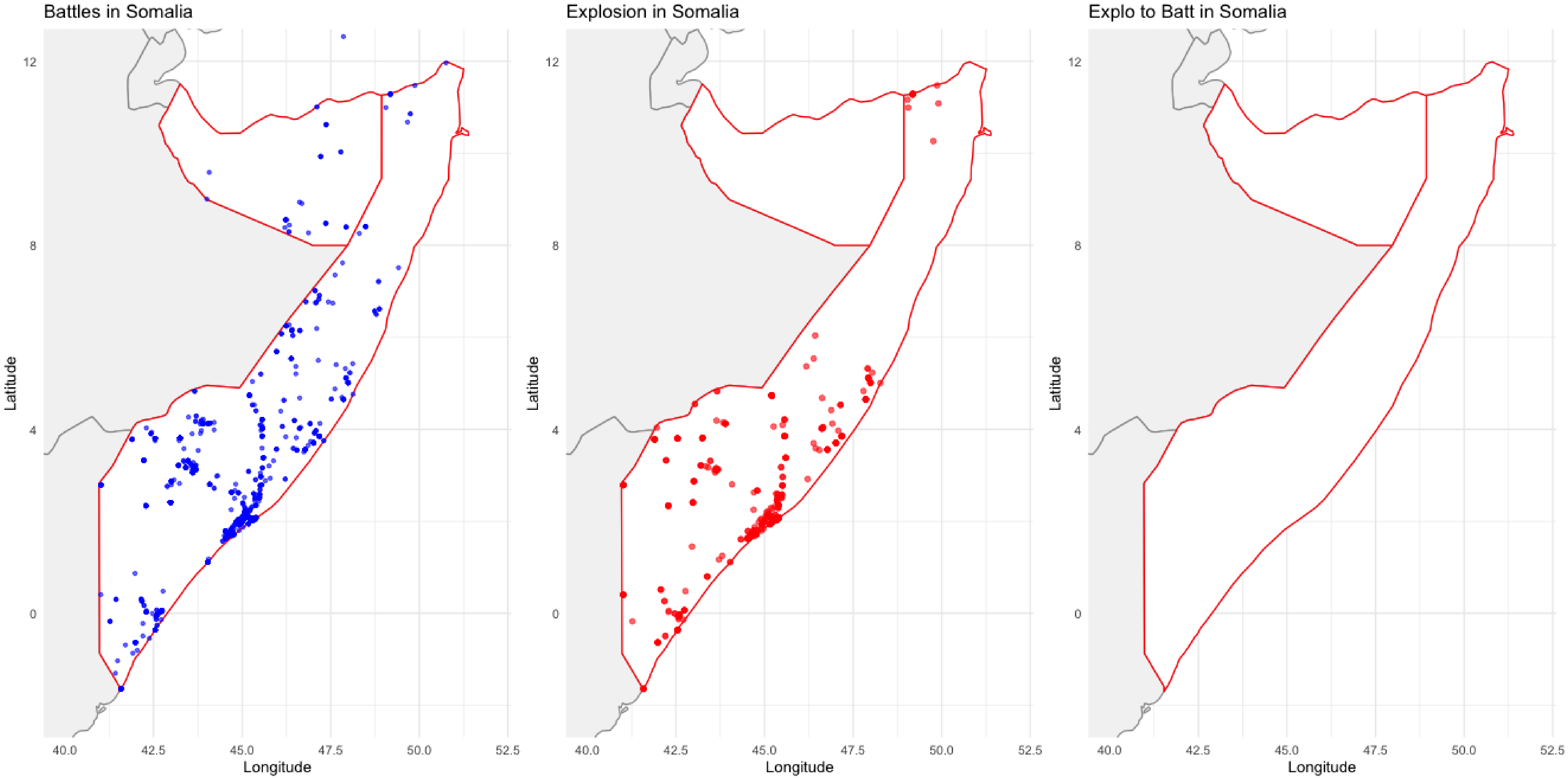
Explo to Batt in Somalia

2023



Battles in Somalia
Explosion in Somalia
Explo to Batt in Somalia
Latitude
12
8
4
0
Longitude
40.0
42.5
45.0
47.5
50.0
52.5

2024



Battles in Somalia

Explosion in Somalia

Explo to Batt in Somalia

- What we can possibly assume from this :
  - There is an interaction between event types (intra), and between events and location. There are some preserve patterns over the years
  - How the "reported" battles and explosion change over the years are proportional. This is seen from how their KM overlaps (even the long tail)
  - Correction does not occur in the short term (we will analyse the delay of correction more thoroughly later). Older events (wrt time of event, not report time) get more chance to be corrected. Maybe the 2024 reported events that we know of now will be corrected in 4 years (something we need to foresee)
  - The older events had more chance to get corrected throughout the years, yet we still observe some correction in 5 months. With more months of observation, we could observe more error correction (the measurement error is then underestimated if we use only the correction during the 5 months)

- What we do not know.
  - The time of report of every event. Which leaves us in doubt because it is also possible that delay of correction is negatively correlated to the delay of report. Which is why we do not have any correction for reported events that happened in 2024 (because they have less than a year of delay report: that at least we know)
  - The error. We only observed the correction in 5 months. But is it enough to say anything about the measurement error given our data exhibit a long tail survival function for delay?

Updates Behaviour

- Can we deduce any useful information from the timestamp of events reported on or before the first download AND did not get updated anymore?

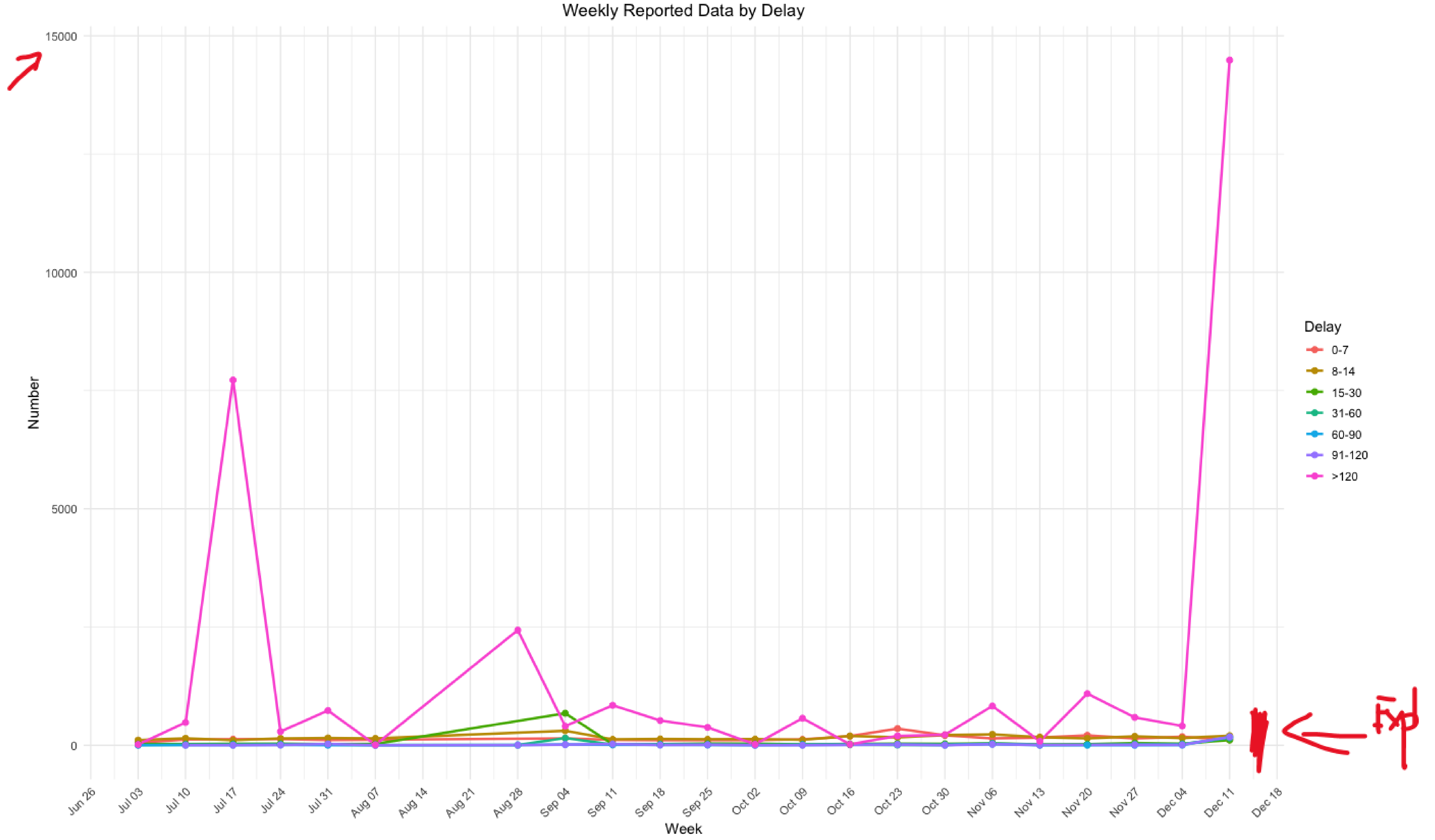
- Problem : The timestamp is ambiguous. It can signify
  - Time of first report
  - Time of last update without error correction
  - Time of last update with error correction

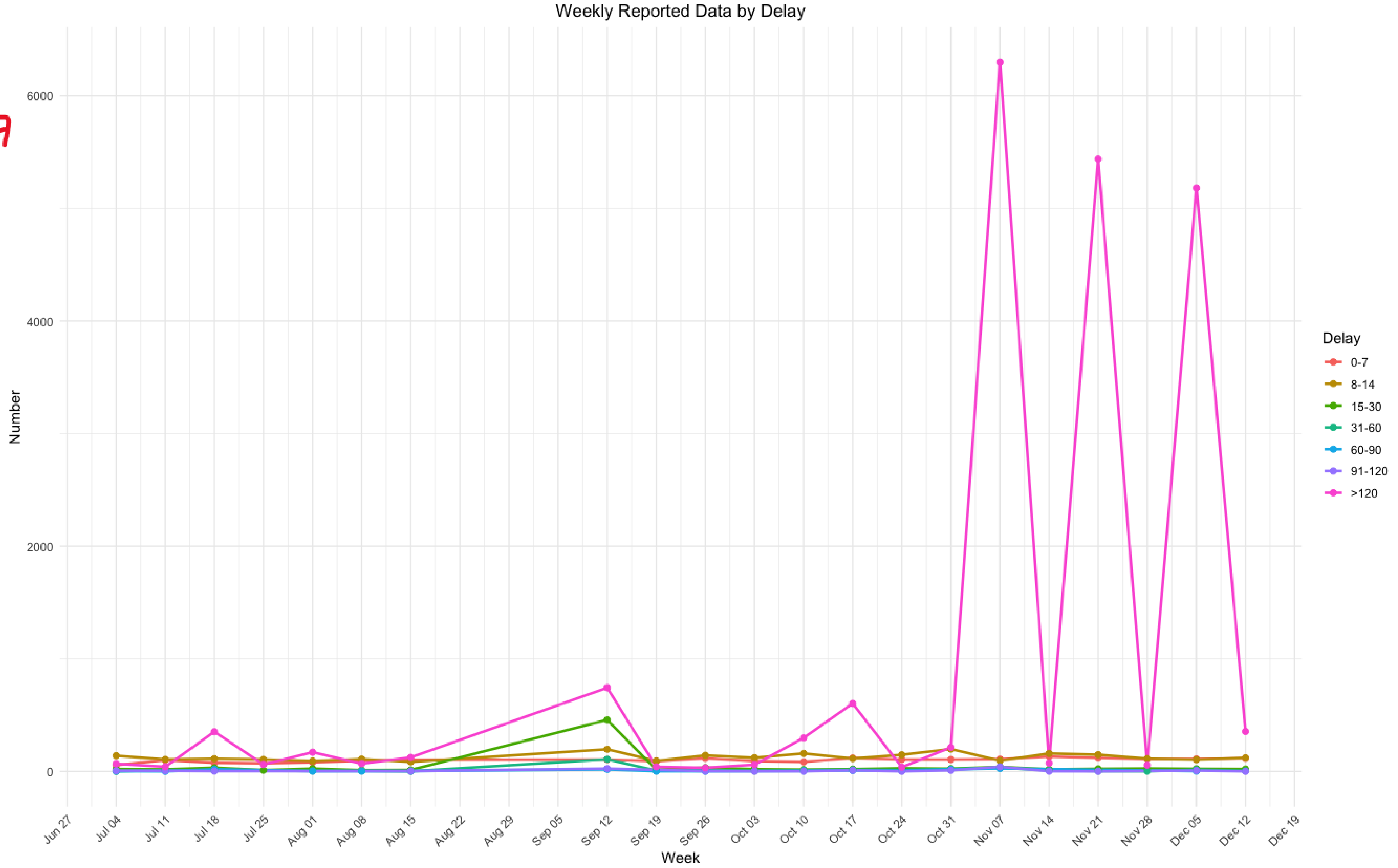
$$T_s = T_e + D_r + \sum_{i=1}^{p} U_i$$

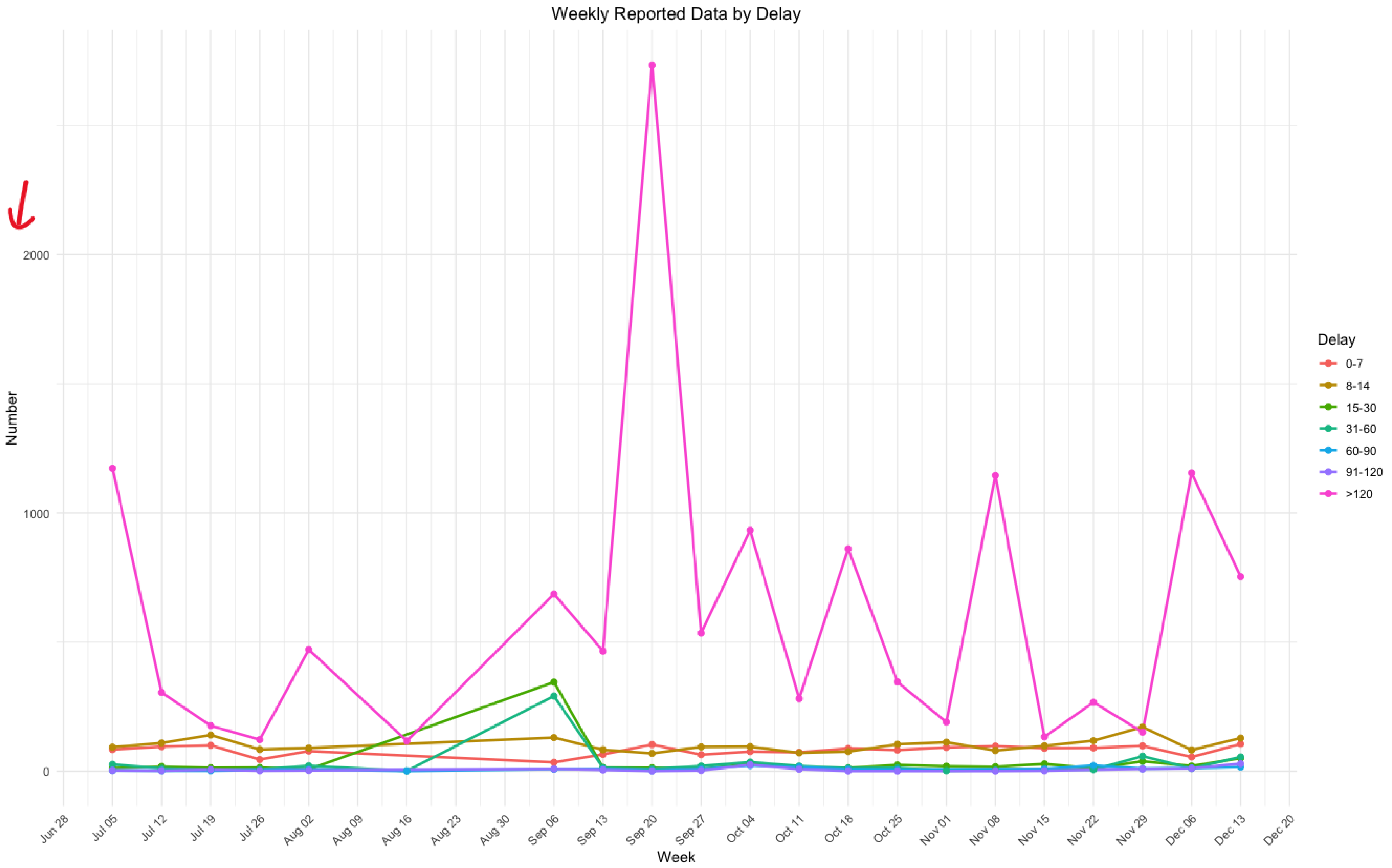
Ts : Last timestamp
Te : Time of event
Dr : Delay of first report
Ui : Delay of i-th update after the (i-1)th
where U_0 = Dr
p : number of updates after first report

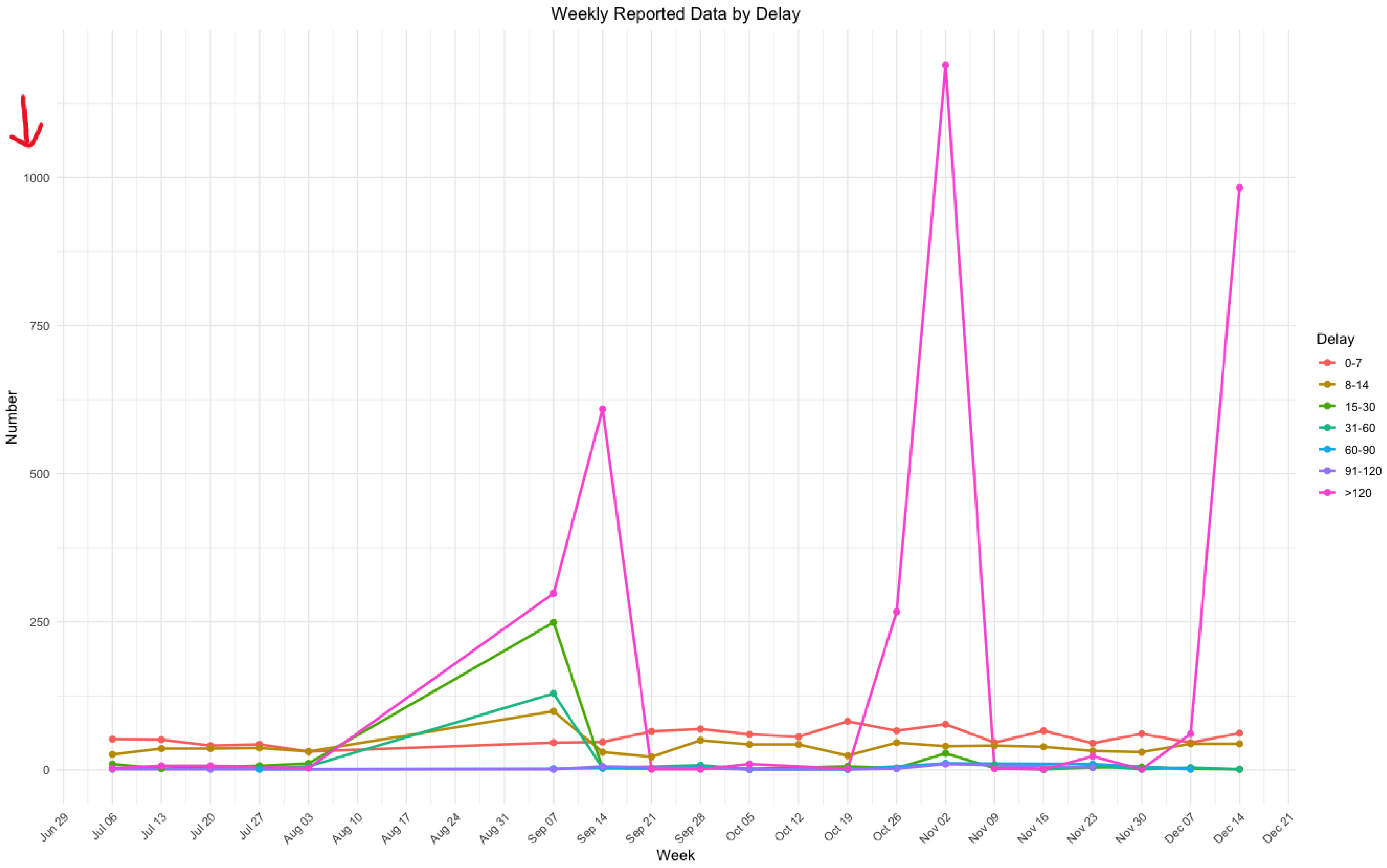
- **Can we then assume that the distribution of Dr and Dr+Sum(U) is the same? (negligeable error correction)** NO

- Assuming delay distribution is **stationary**, the KM delays of the events newly reported between our first dowload and the followings AND the KM delays of the events reported during the same period the year before by using the difference between time of event and timestamp **should not have significant difference** (just with errors due to randomness) WRONG


Weekly Reported Data by Delay
15000
10000
5000
0
Number
Week
Jun 26
Jul 03
Jul 10
Jul 17
Jul 24
Jul 31
Aug 07
Aug 14
Aug 21
Aug 28
Sep 04
Sep 11
Sep 18
Sep 25
Oct 02
Oct 09
Oct 16
Oct 23
Oct 30
Nov 06
Nov 13
Nov 20
Nov 27
Dec 04
Dec 11
Dec 18
Delay
0-7
8-14
15-30
31-60
60-90
91-120
>120
Exp


Weekly Reported Data by Delay
Number
6000
4000
2000
0
Week
Jun 27
Jul 04
Jul 11
Jul 18
Jul 25
Aug 01
Aug 08
Aug 15
Aug 22
Aug 29
Sep 05
Sep 12
Sep 19
Sep 26
Oct 03
Oct 10
Oct 17
Oct 24
Oct 31
Nov 07
Nov 14
Nov 21
Nov 28
Dec 05
Dec 12
Dec 19
Delay
0-7
8-14
15-30
31-60
60-90
91-120
>120


Weekly Reported Data by Delay
Number
2000
1000
0
Week
Jun 28
Jul 05
Jul 12
Jul 19
Jul 26
Aug 02
Aug 09
Aug 16
Aug 23
Aug 30
Sep 06
Sep 13
Sep 20
Sep 27
Oct 04
Oct 11
Oct 18
Oct 25
Nov 01
Nov 08
Nov 15
Nov 22
Nov 29
Dec 06
Dec 13
Dec 20
Delay
0-7
8-14
15-30
31-60
60-90
91-120
>120

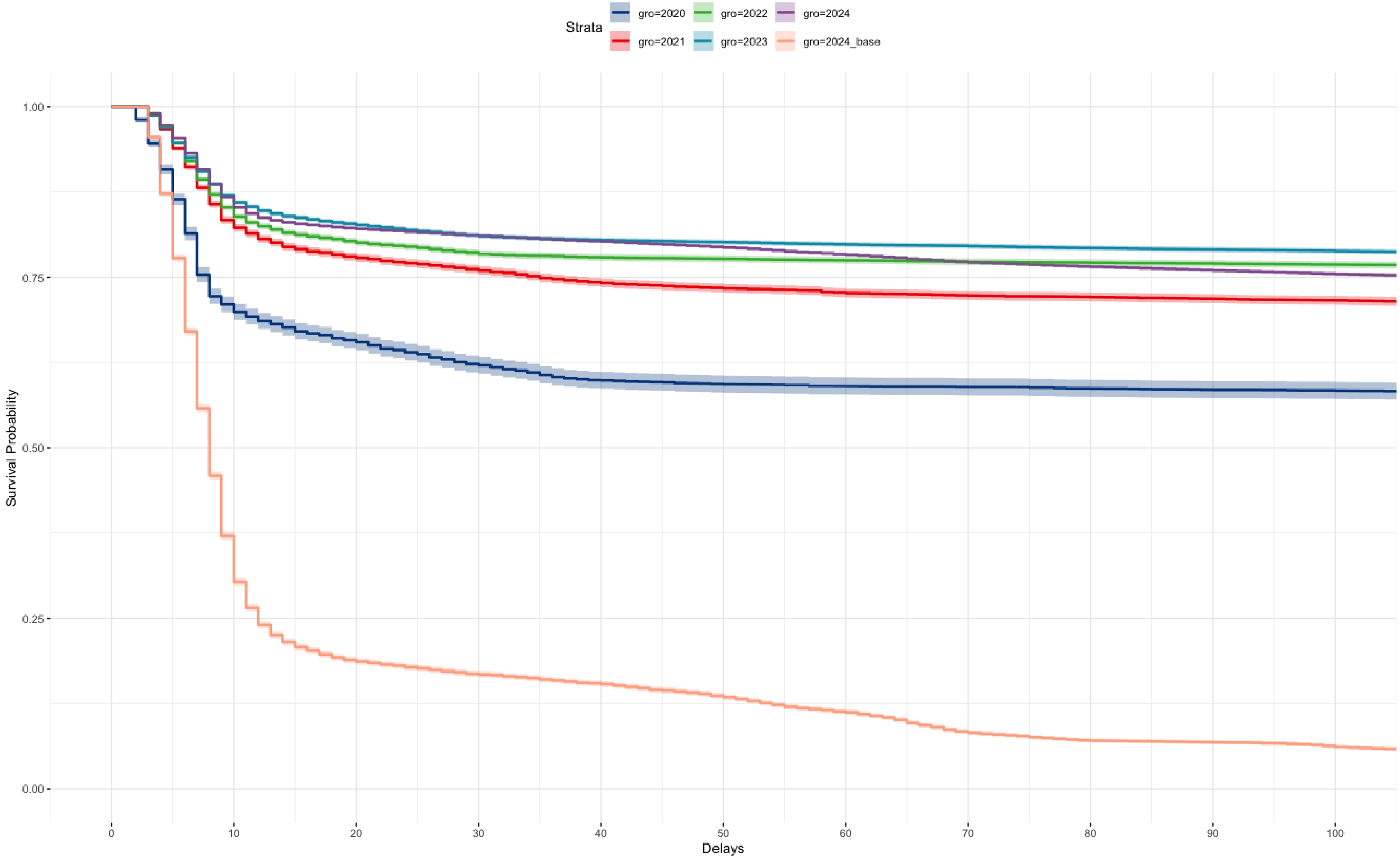

Weekly Reported Data by Delay
Number
1000
750
500
250
0
Week
Jun 29
Jul 06
Jul 13
Jul 20
Jul 27
Aug 03
Aug 10
Aug 17
Aug 24
Aug 31
Sep 07
Sep 14
Sep 21
Sep 28
Oct 05
Oct 12
Oct 19
Oct 26
Nov 02
Nov 09
Nov 16
Nov 23
Nov 30
Dec 07
Dec 14
Dec 21
Delay
0-7
8-14
15-30
31-60
60-90
91-120
>120

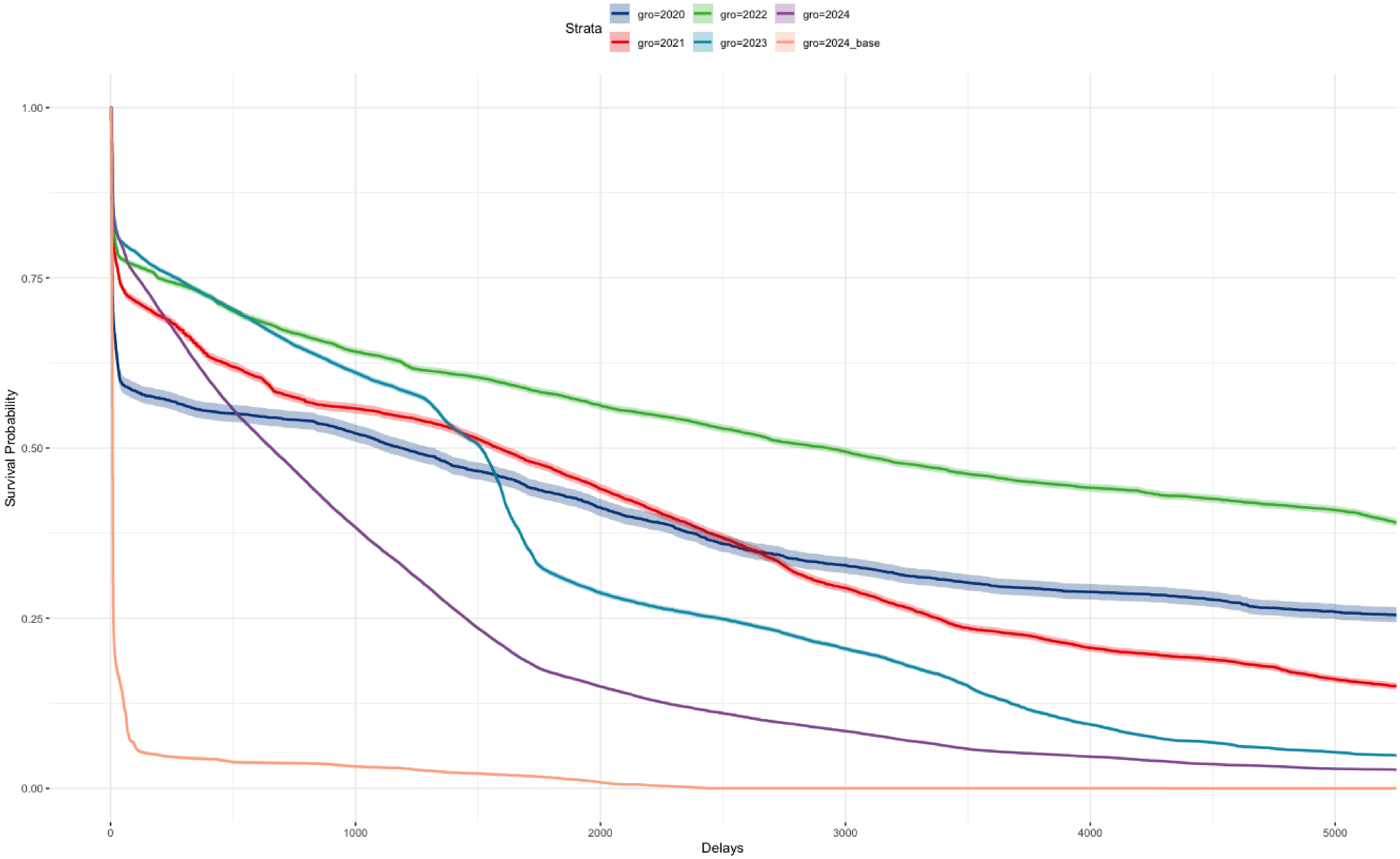
Discussion

- It's either we have consistently relatively high number of late reports or late updates or both (compared to the event time)

- The number of events with less delay of report and cumulated delay of update is very small. These indicate the events that give us more evidence that they didn't get updated (because Dr+Sum(U) is small). The inverse is not true. The events with high Dr+Sum(U) could be events with high Dr or high Sum(U) or both

- The number of events reported/updated decreases every year. A rational explanation is that there is a migration from preceding to the next years. Hence we have many updates. And the older the event is the higher is the probability of it being updated (or corrected)


Strata
gro=2020
gro=2021
gro=2022
gro=2023
gro=2024
gro=2024_base
Survival Probability
1.00
0.75
0.50
0.25
0.00
Delays
0
10
20
30
40
50
60
70
80
90
100


Strata
gro=2020
gro=2022
gro=2024
gro=2021
gro=2023
gro=2024_base
Survival Probability
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
5000
Delays

$$S_{2023}(t) \geq S_{2022}(t) \geq S_{2021}(t) \geq S_{2020}(t)$$

In the short delay

We see big difference between KP 2024 (where we took only new events and no update) and the others where we took the last date of update.

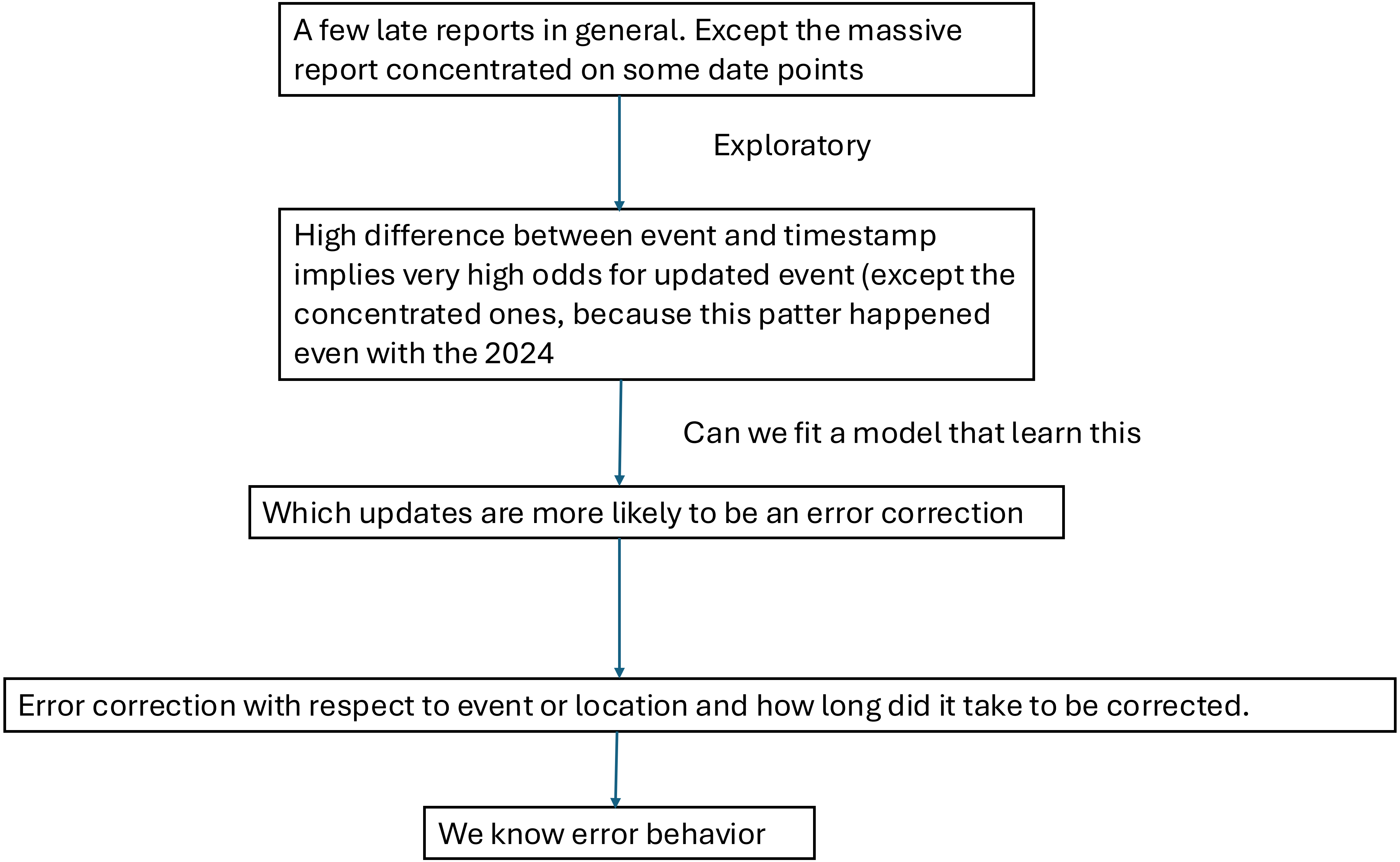
Hence the update is then very important because many of the reported events will be updated.

Another discussion

- We talked about update but not every update is important enough to be considered "error". However we could prove with common sense that update is underestimated using only updates observed during the 5 months. The latter increases the evidence that our error is also underestimated (with the assumption that error is proportional to the updated).

- How can we then estimate the "relevant" part of the update. Our best shot is to analyse these changes that occurred in the 5 months of observation and find a pattern.

- If we have date of event, covariates, last update, can we tell with some level of confidence if it was an event that got corrected? And got corrected from what (ex : Explo to Battle?) and when did it get corrected?



A few late reports in general. Except the massive report concentrated on some date points

Exploratory

High difference between event and timestamp implies very high odds for updated event (except the concentrated ones, because this patter happened even with the 2024

Can we fit a model that learn this

Which updates are more likely to be an error correction

Error correction with respect to event or location and how long did it take to be corrected.

We know error behavior