INTERACT | SenseDoc Quality checks

B. Thierry, Spherelab

25 September, 2025

Contents

1 QA objectives

Create summary statistics (see `QA` subfolder):

- date ranges
- number of days of data per participant
- min, max, SD distributions
- GPS locations
- [Sept. 2025] Step statistics

2 Data coverage

2.1 Get data from database

```
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'mtl' city_id, 1 wave_id
        ,ts.*, tm.wearing
    FROM top_sd.top_1sec_mtl ts,
        top_sd.top_1min_mtl tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
```

```
  -- Get Mtl / w2
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'mtl' city_id, 2 wave_id
        ,ts.*, tm.wearing
    FROM top_sd2.top_1sec_mtl ts,
        top_sd2.top_1min_mtl tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
  -- Get Mtl / w3
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'mtl' city_id, 3 wave_id
        ,ts.*, tm.wearing
    FROM top_sd3.top_1sec_mtl ts,
        top_sd3.top_1min_mtl tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
  -- Get Skt / w1
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'skt' city_id, 1 wave_id
        ,ts.*, tm.wearing
    FROM top_sd.top_1sec_skt ts,
        top_sd.top_1min_skt tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
  -- Get Skt / w2
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
```

```
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'skt' city_id, 2 wave_id
        ,ts.*, tm.wearing
    FROM top_sd2.top_1sec_skt ts,
        top_sd2.top_1min_skt tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
  -- Get skt / w3
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'skt' city_id, 3 wave_id
        ,ts.*, tm.wearing
    FROM top_sd3.top_1sec_skt ts,
        top_sd3.top_1min_skt tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
  -- Get van / w1
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'van' city_id, 1 wave_id
        ,ts.*, tm.wearing
    FROM top_sd.top_1sec_van ts,
        top_sd.top_1min_van tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
  -- Get van / w2
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'van' city_id, 2 wave_id
```

```
        ,ts.*, tm.wearing
    FROM top_sd2.top_1sec_van ts,
        top_sd2.top_1min_van tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
  -- Get van / w3
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'van' city_id, 3 wave_id
        ,ts.*, tm.wearing
    FROM top_sd3.top_1sec_van ts,
        top_sd3.top_1min_van tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
  -- Get vic / w1
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'vic' city_id, 1 wave_id
        ,ts.*, tm.wearing
    FROM top_sd.top_1sec_vic ts,
        top_sd.top_1min_vic tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
  -- Get vic / w2
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'vic' city_id, 2 wave_id
        ,ts.*, tm.wearing
    FROM top_sd2.top_1sec_vic ts,
        top_sd2.top_1min_vic tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
```

```
group by city_id, wave_id, interact_id, sd_id
-- Get vic / w3
UNION
select city_id, wave_id, interact_id, sd_id
    ,min(utcdate) start_time, max(utcdate) end_time
    ,max(utcdate) - min(utcdate) survey_duration
    ,count(*) n_epoch
    ,sum(wearing) n_epoch_wearing
    ,count(lat) n_gps_fix
from (
    SELECT 'vic' city_id, 3 wave_id
        ,ts.*, tm.wearing
    FROM top_sd3.top_1sec_vic ts,
        top_sd3.top_1min_vic tm
    WHERE ts.interact_id = tm.interact_id AND ts.sd_id = tm.sd_id
        AND date_trunc('minute', ts.utcdate) = tm.utcdate) as foo
group by city_id, wave_id, interact_id, sd_id
```

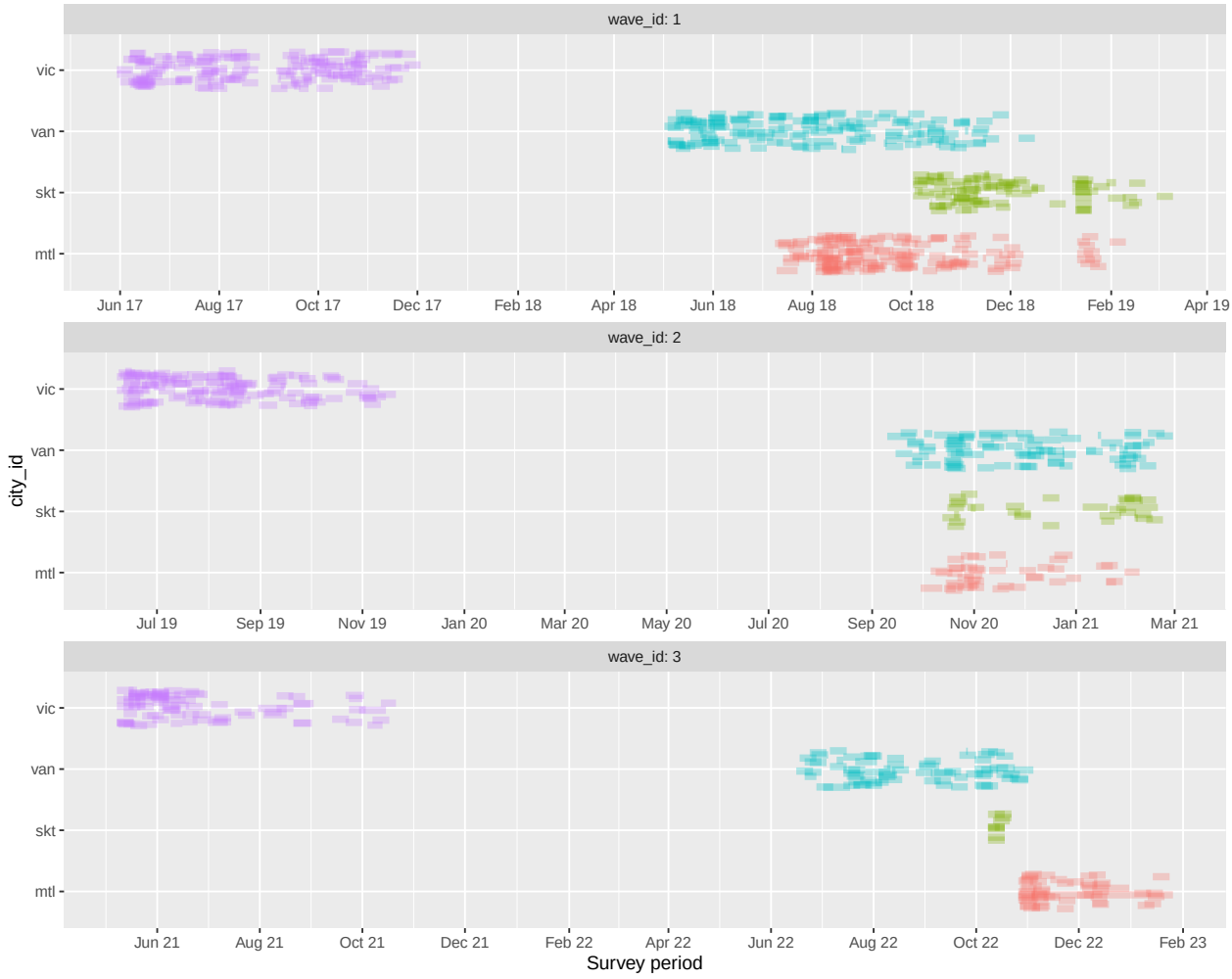
```
head(top_1s_agg)
```

city_id	wave_id	interact_id	sd_id	start_time	end_time	survey_duration	n_epoch	n_epoch_wearing	n_gps_fix
vic	3	103772760	357	2021-05-25 16:08:19	2021-06-04 06:59:58	9 days 14:51:39	831100	550140	59853
skt	2	302619633	312	2020-10-20 13:09:32	2020-10-23 14:58:38	3 days 01:49:06	265747	145719	15200
van	3	203842375	451	2022-07-24 13:31:40	2022-08-03 06:13:11	9 days 16:41:31	837692	589752	77296
van	1	201585258	445	2018-08-22 14:07:11	2018-09-01 03:12:32	9 days 13:05:21	824722	431133	261061
mtl	1	401178032	96	2019-01-11 11:41:22	2019-01-21 02:10:44	9 days 14:29:22	829763	524205	275790
van	2	201140482	375	2020-09-10 14:59:52	2020-09-27 03:35:11	16 days 12:35:19	1427720	455100	223106

2.2 Survey time span

```
top_1s_agg |>
  mutate(
    sdate = as_date(start_time),
    edate = as_date(end_time)
  ) |>
  ggplot() +
  geom_segment(
    aes(
      x = sdate,
      xend = edate,
      y = city_id,
      color = city_id
  ),
    linewidth = 2,
    alpha = .3,
    position = position_jitter(width = 0, height = .3)
```

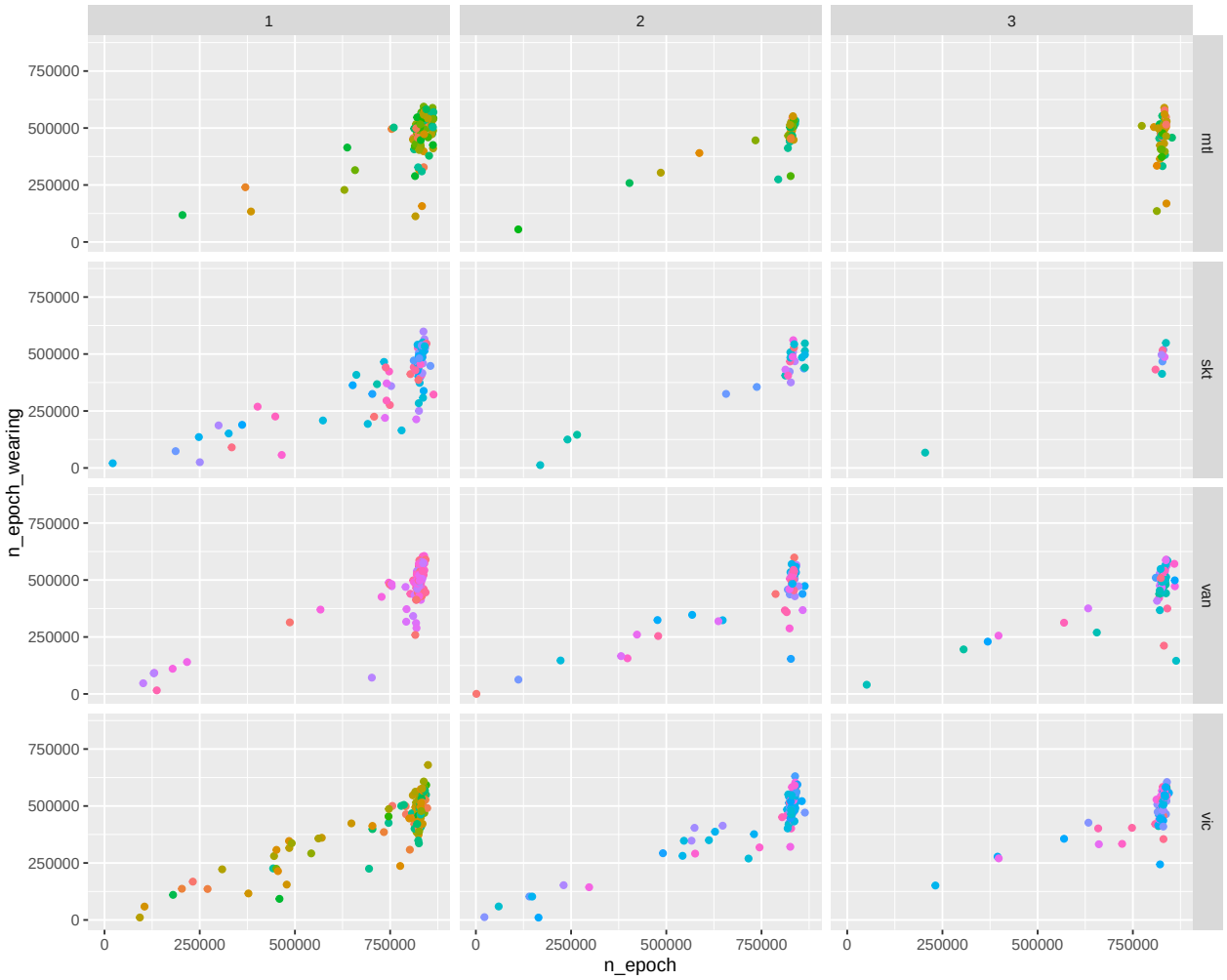
```
  ) +
  scale_x_date(name = "Survey period", date_breaks = "2 month", date_labels = "%b %y") +
  facet_wrap(vars(wave_id), ncol = 1, scales = "free", labeller = label_both) +
  theme(legend.position = "none")
```



2.3 Wear time vs total survey time

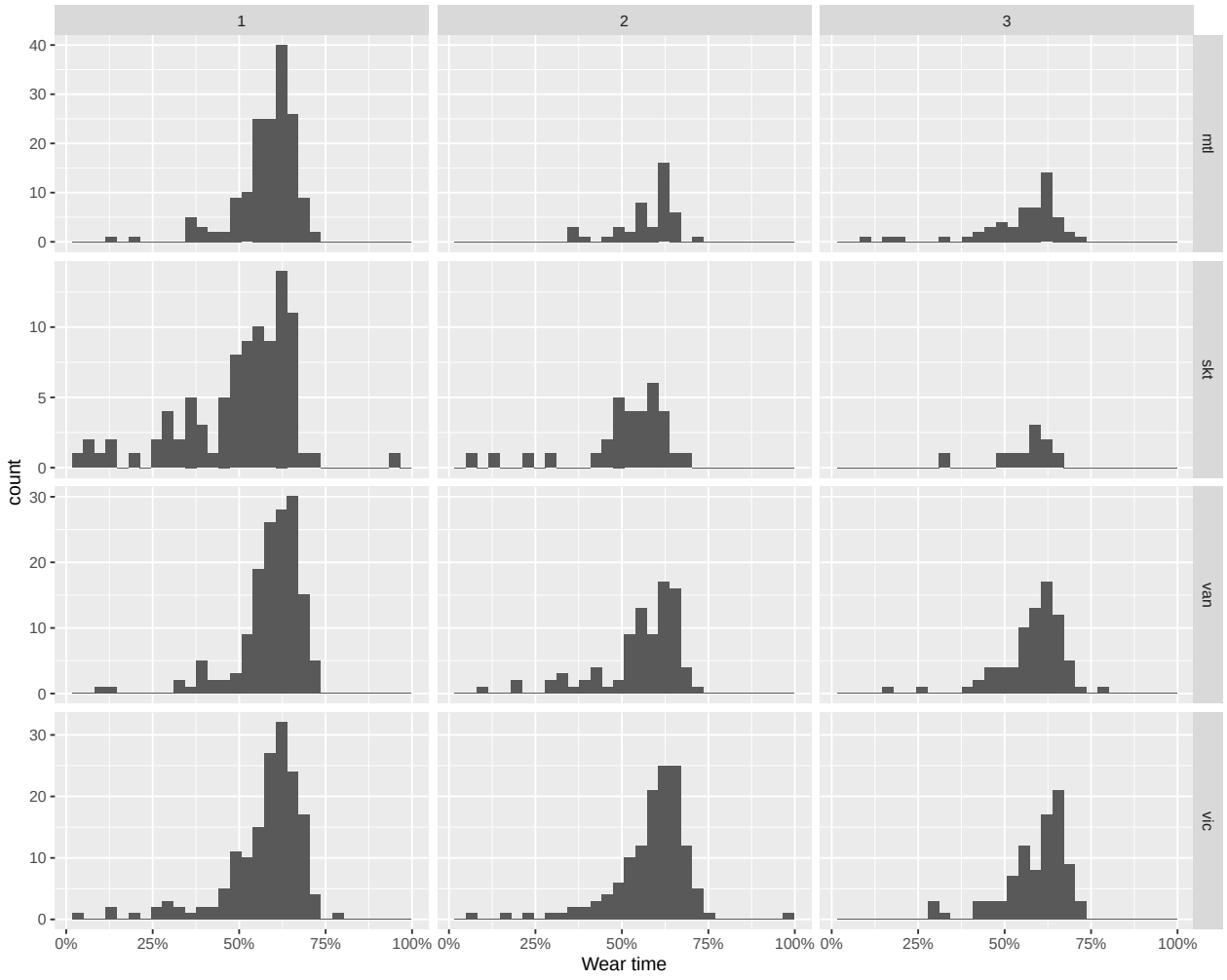
```
top_1s_agg |>
  ggplot() +
  geom_point(aes(x=n_epoch, y=n_epoch_wearing, color=factor(sd_id))) +
  xlim(0, 10 * 24 * 3600) + ylim(0, 10 * 24 * 3600) + # Define theoretical max of survey n_epoch, i.e.
  facet_grid(rows = vars(city_id), cols = vars(wave_id)) +
  theme(legend.position = "none")
```

```
## Warning: Removed 156 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
top_1s_agg |>
  mutate(`Wear time` = n_epoch_wearing / n_epoch) |>
  ggplot() +
  geom_histogram(aes(x = `Wear time`)) +
  scale_x_continuous(labels = scales::percent) +
  facet_grid(rows = vars(city_id), cols = vars(wave_id), scales = "free_y")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



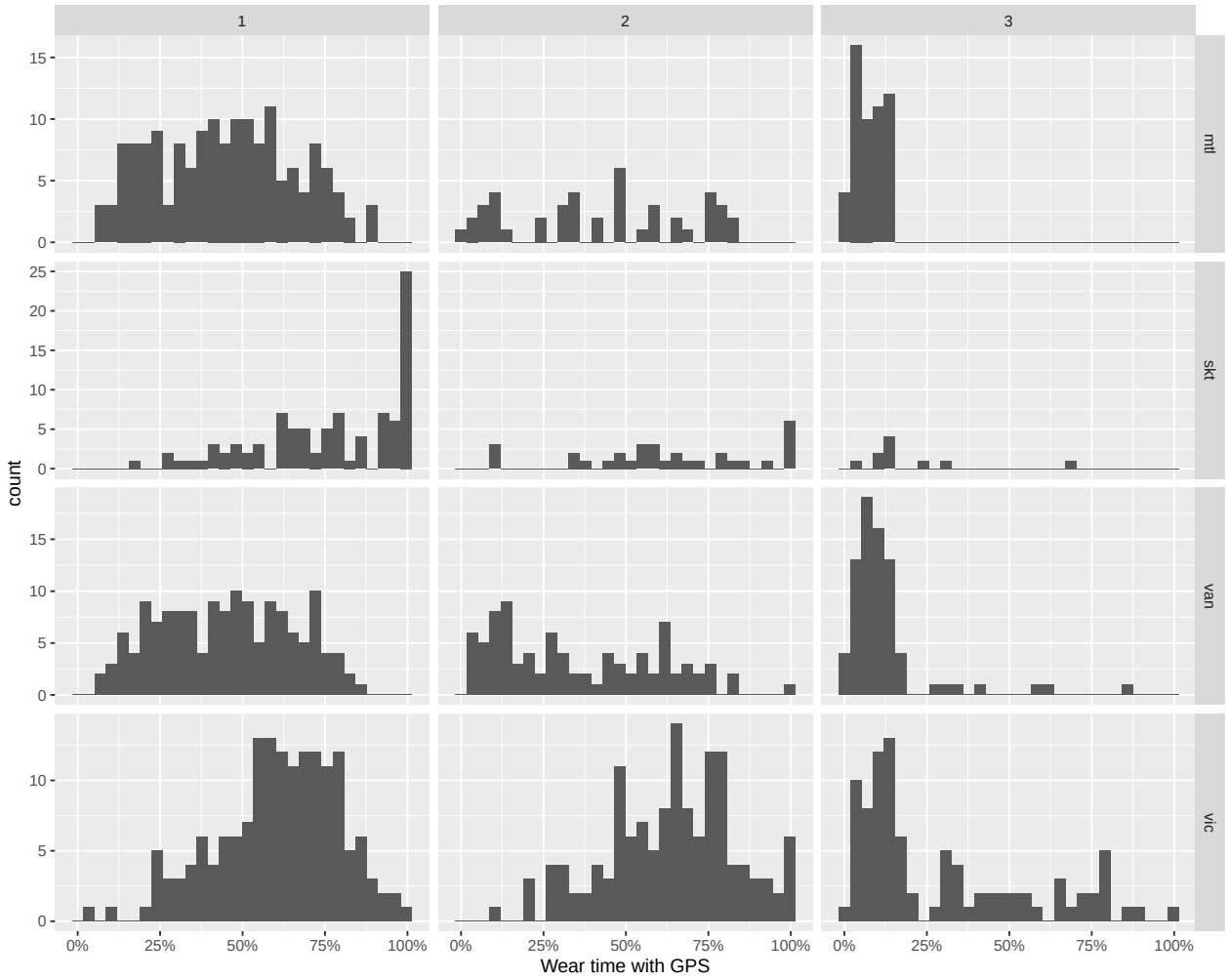
2.4 Wear time with GPS fix

```
top_1s_agg |>
  ggplot() +
  geom_point(aes(x=n_epoch_wearing, y=n_gps_fix, color=factor(sd_id))) +
  xlim(0, 10 * 24 * 3600) + ylim(0, 10 * 24 * 3600) + # Define theoretical max of survey n_epoch, i.e.
  facet_grid(rows = vars(city_id), cols = vars(wave_id)) +
  theme(legend.position = "none")
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```


```
top_1s_agg |>
  mutate(`Wear time with GPS` = pmin(n_gps_fix / n_epoch_wearing, 1)) |>
  ggplot() +
  geom_histogram(aes(x = `Wear time with GPS`)) +
  scale_x_continuous(labels = scales::percent) +
  facet_grid(rows = vars(city_id), cols = vars(wave_id), scales = "free_y")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2.5 Step statistcs

Steps have been added in September 2025 following the Bernard Asante's work. Steps are computing according to the Python package `stepcount` (see repo) based on *Small SR, Chan S, Walmsley R, et al. (2024) Self-Supervised Machine Learning to Characterize Step Counts from Wrist-Worn Accelerometers in the UK Biobank. Medicine & Science in Sports & Exercise. DOI: 10.1249/MSS.0000000000003478*

```
select 'mtl' city_id, 1 wave_id,
  utcdate at time zone 'America/Montreal' datetime,
  wearing, steps, steps_adj
from top_sd.top_1min_mtl
union
select 'skt' city_id, 1 wave_id,
  utcdate at time zone 'America/Regina' datetime,
  wearing, steps, steps_adj
from top_sd.top_1min_skt
union
select 'van' city_id, 1 wave_id,
  utcdate at time zone 'America/Vancouver' datetime,
  wearing, steps, steps_adj
from top_sd.top_1min_van
union
```

```
select 'vic' city_id, 1 wave_id,
  utcdate at time zone 'America/Vancouver' datetime,
  wearing, steps, steps_adj
from top_sd.top_1min_vic
union
select 'mtl' city_id, 2 wave_id,
  utcdate at time zone 'America/Montreal' datetime,
  wearing, steps, steps_adj
from top_sd2.top_1min_mtl
union
select 'skt' city_id, 2 wave_id,
  utcdate at time zone 'America/Regina' datetime,
  wearing, steps, steps_adj
from top_sd2.top_1min_skt
union
select 'van' city_id, 2 wave_id,
  utcdate at time zone 'America/Vancouver' datetime,
  wearing, steps, steps_adj
from top_sd2.top_1min_van
union
select 'vic' city_id, 2 wave_id,
  utcdate at time zone 'America/Vancouver' datetime,
  wearing, steps, steps_adj
from top_sd.top_1min_vic
union
select 'mtl' city_id, 3 wave_id,
  utcdate at time zone 'America/Montreal' datetime,
  wearing, steps, steps_adj
from top_sd3.top_1min_mtl
union
select 'skt' city_id, 3 wave_id,
  utcdate at time zone 'America/Regina' datetime,
  wearing, steps, steps_adj
from top_sd3.top_1min_skt
union
select 'van' city_id, 3 wave_id,
  utcdate at time zone 'America/Vancouver' datetime,
  wearing, steps, steps_adj
from top_sd3.top_1min_van
union
select 'vic' city_id, 3 wave_id,
  utcdate at time zone 'America/Vancouver' datetime,
  wearing, steps, steps_adj
from top_sd3.top_1min_vic
```

2.5.1 All epochs

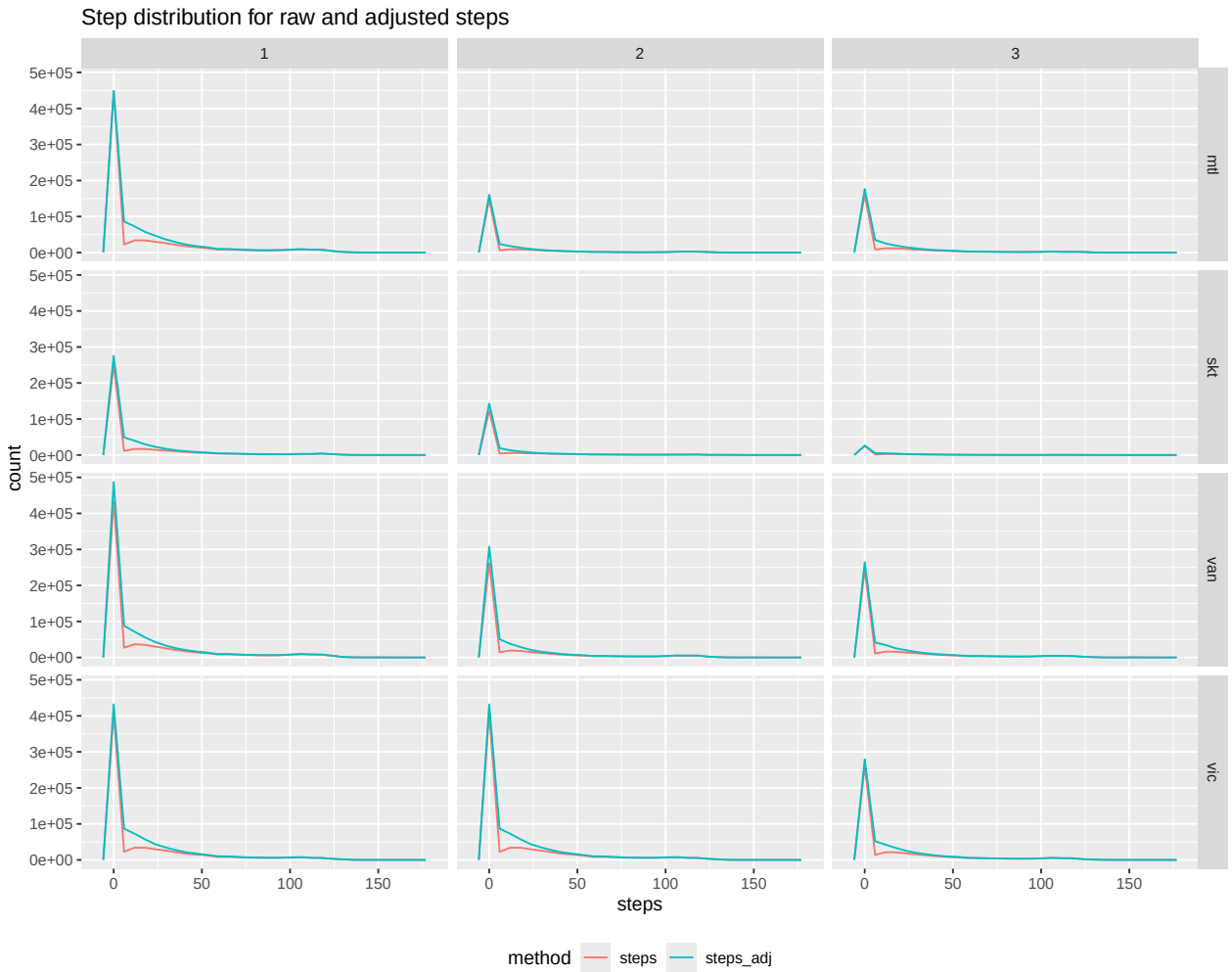
```
top_1m |>
  select(steps, steps_adj) |>
  summary()
```

```
##      steps           steps_adj
##  Min.   :  0       Min.   :  0.00
```

```
##  1st Qu.:  0       1st Qu.:  0.00
##  Median :  0       Median :  1.00
##  Mean   : 17       Mean   : 16.29
##  3rd Qu.: 23       3rd Qu.: 21.00
##  Max.   :171       Max.   :170.00
##  NA's   :4305757   NA's   :2866640
```

```
top_1m |>
  pivot_longer(cols = starts_with("steps"), names_to = "method", values_to = "steps", values_drop_na = T
  ggplot() +
  geom_freqpoly(aes(x = steps, color = method)) +
  facet_grid(rows = vars(city_id), cols = vars(wave_id)) +
  labs(title = "Step distribution for raw and adjusted steps") +
  theme(legend.position = "bottom")
```

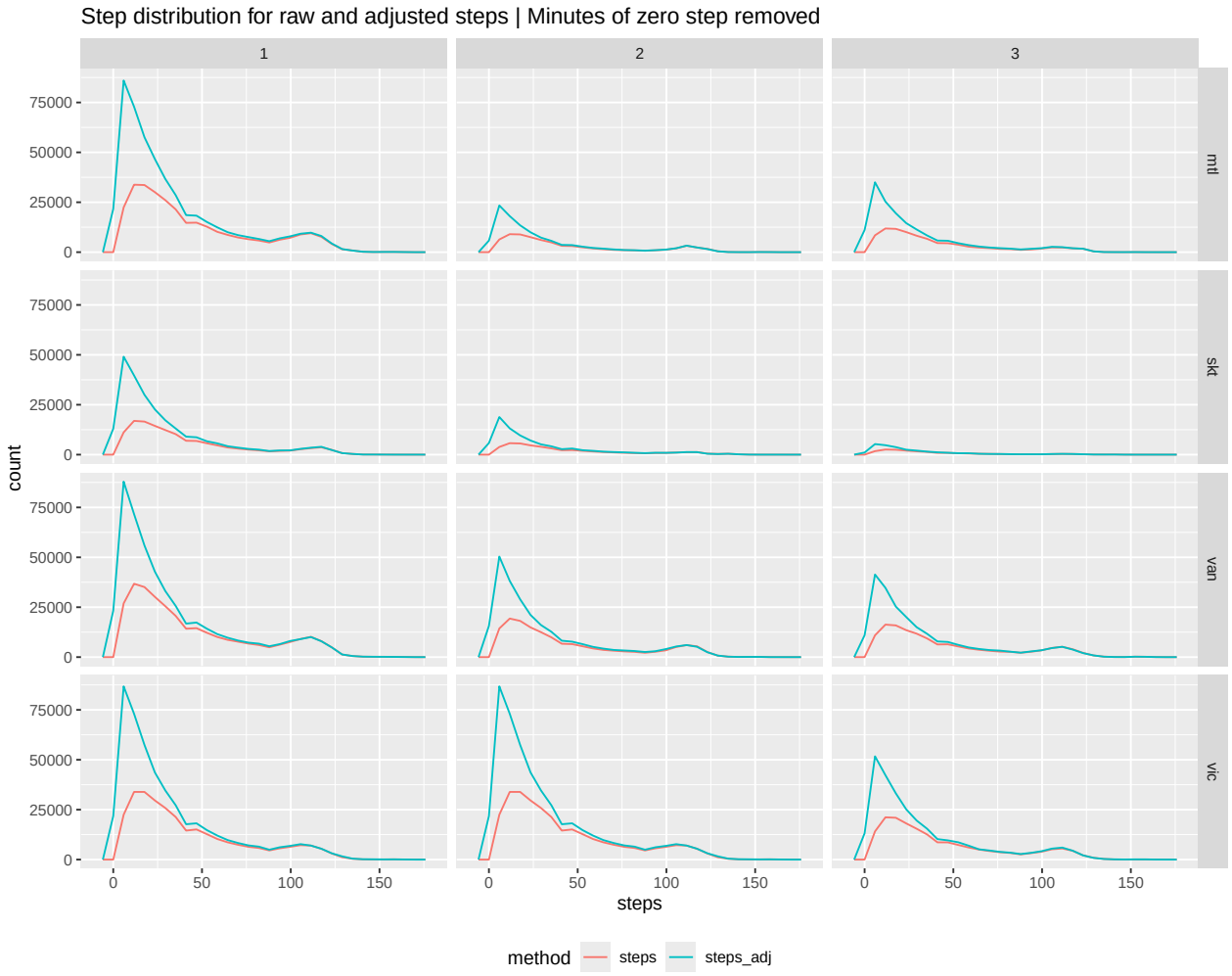
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
top_1m |>
  pivot_longer(cols = starts_with("steps"), names_to = "method", values_to = "steps", values_drop_na = T
  filter(steps != 0) |>
  ggplot() +
  geom_freqpoly(aes(x = steps, color = method)) +
  facet_grid(rows = vars(city_id), cols = vars(wave_id)) +
```

```
  labs(title = "Step distribution for raw and adjusted steps | Minutes of zero step removed") +
  theme(legend.position = "bottom")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



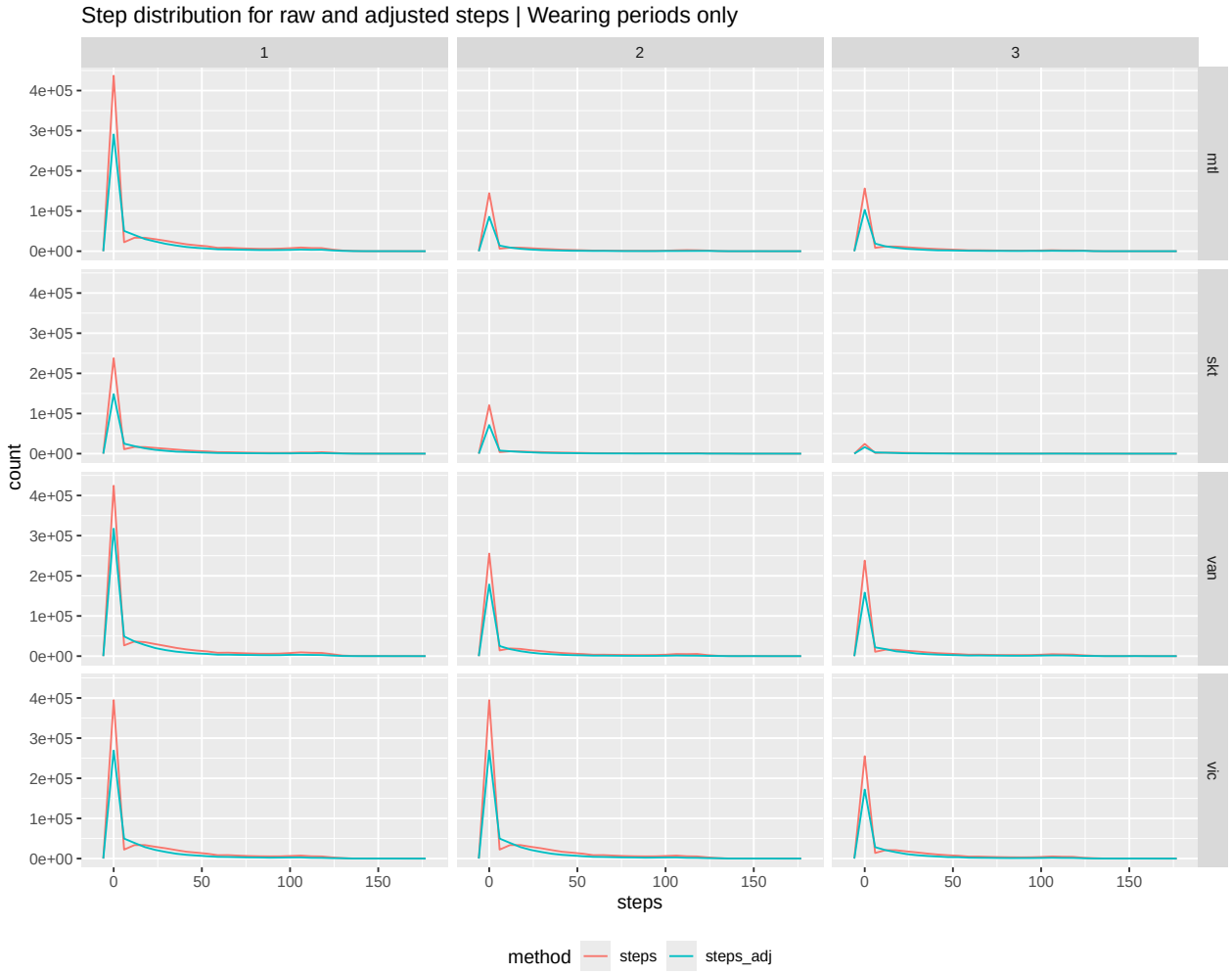
2.5.2 Wearing period epochs

```
top_1m |>
  filter(wearing == 1) |>
  select(steps, steps_adj) |>
  summary()
```

```
##      steps          steps_adj
##  Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  0.00   1st Qu.:  0.00
##  Median :  0.00   Median :  0.00
##  Mean   : 17.19   Mean   : 12.38
##  3rd Qu.: 23.00   3rd Qu.: 14.00
##  Max.   :171.00   Max.   :170.00
##  NA's   :424653   NA's   :1952965
```

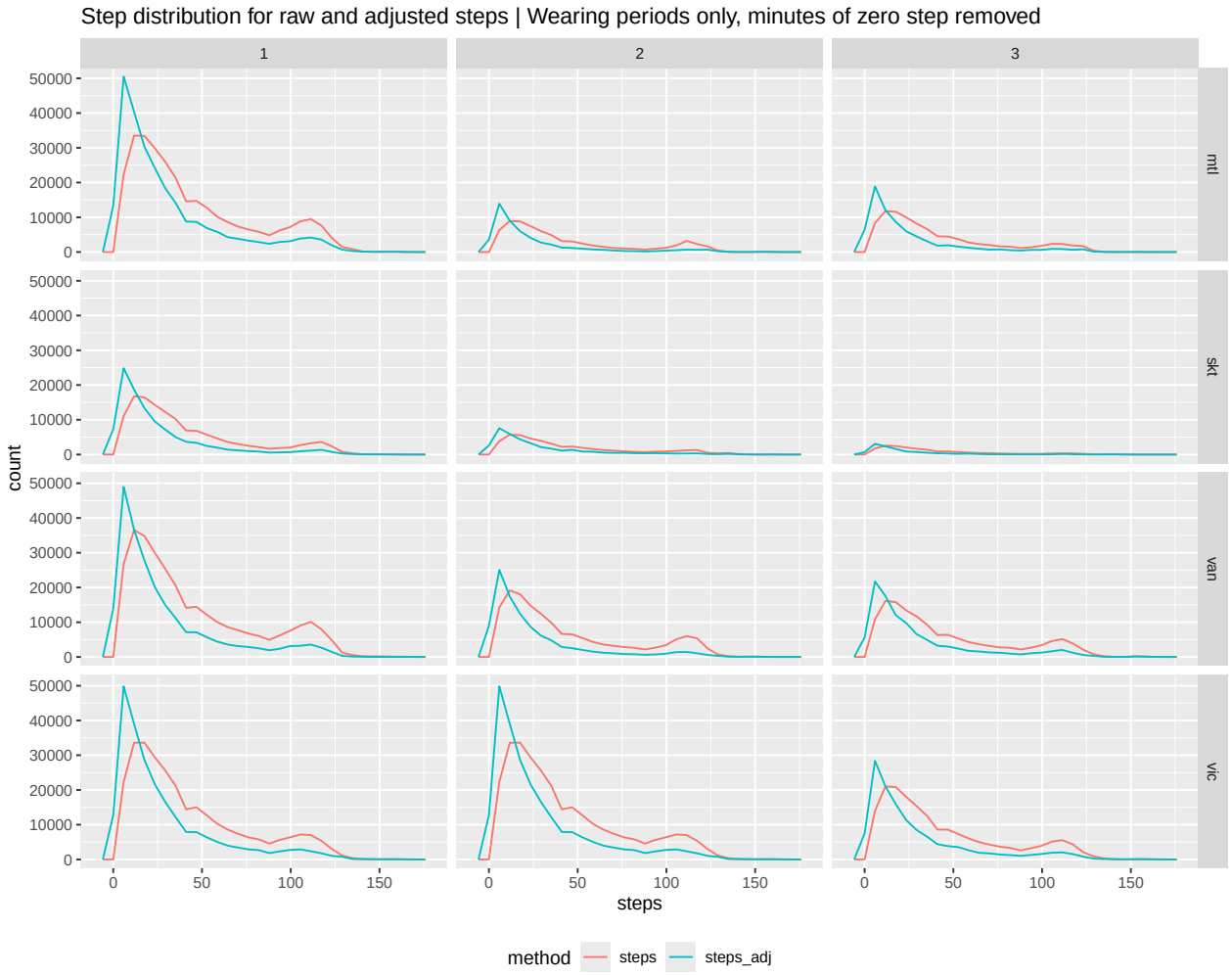
```
top_1m |>
  pivot_longer(cols = starts_with("steps"), names_to = "method", values_to = "steps", values_drop_na = T
  filter(wearing == 1) |>
  ggplot() +
  geom_freqpoly(aes(x = steps, color = method)) +
  facet_grid(rows = vars(city_id), cols = vars(wave_id)) +
  labs(title = "Step distribution for raw and adjusted steps | Wearing periods only") +
  theme(legend.position = "bottom")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
top_1m |>
  pivot_longer(cols = starts_with("steps"), names_to = "method", values_to = "steps", values_drop_na = T
  filter(wearing == 1 & steps != 0) |>
  ggplot() +
  geom_freqpoly(aes(x = steps, color = method)) +
  facet_grid(rows = vars(city_id), cols = vars(wave_id)) +
  labs(title = "Step distribution for raw and adjusted steps | Wearing periods only, minutes of zero st
  theme(legend.position = "bottom")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Step distribution for raw and adjusted steps | Wearing periods only, minutes of zero step removed
1
2
3
mtl
skt
van
vic
count
steps
method
steps
steps_adj