Stat 565

Trend And Decompositions

Jan 12 2016

Charlotte Wickham

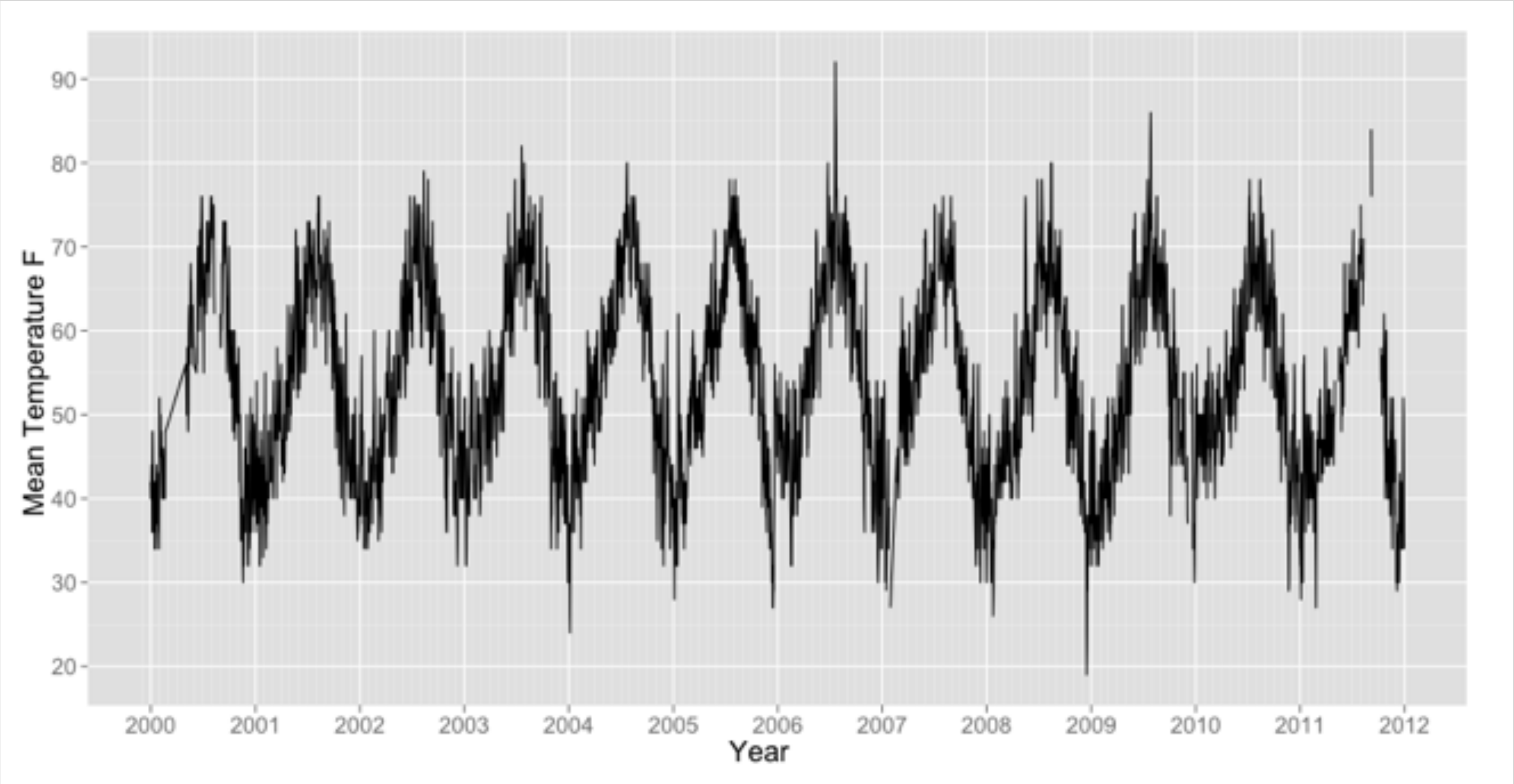
stat565.cwick.co.nz

Announcements

Tomorrow only:

Office hour 2-3pm not 1-2pm

Your turn



Brainstorm: How could we get a feel for the long term trend in this series?

Exploring the trend

It's often hard to see a trend in the presence of noise or seasonality, some options:

Aggregate

use annual data instead

Smooth

moving average or add a line or curve

Subtract

remove noise/seasonality

Aggregate

To examine the long term trend we could average out the seasonality by averaging over the length of the seasonal cycle.

E.g. with temperature, average over a year.

We saw how to do this with dplyr, last week.

We could also achieve this same goal using a model.

Aggregate

We could fit a regression model with a fixed effect for each year, and examine those fixed effects,

$$\text{temp}_t = \mu_{\text{year}(t)} + \text{noise}$$

different mean
for each year

```
year_fit <- lm(temp ~ factor(year), data = corv,
               na.action = na.exclude)
```

I want to treat year
like a category
not a number

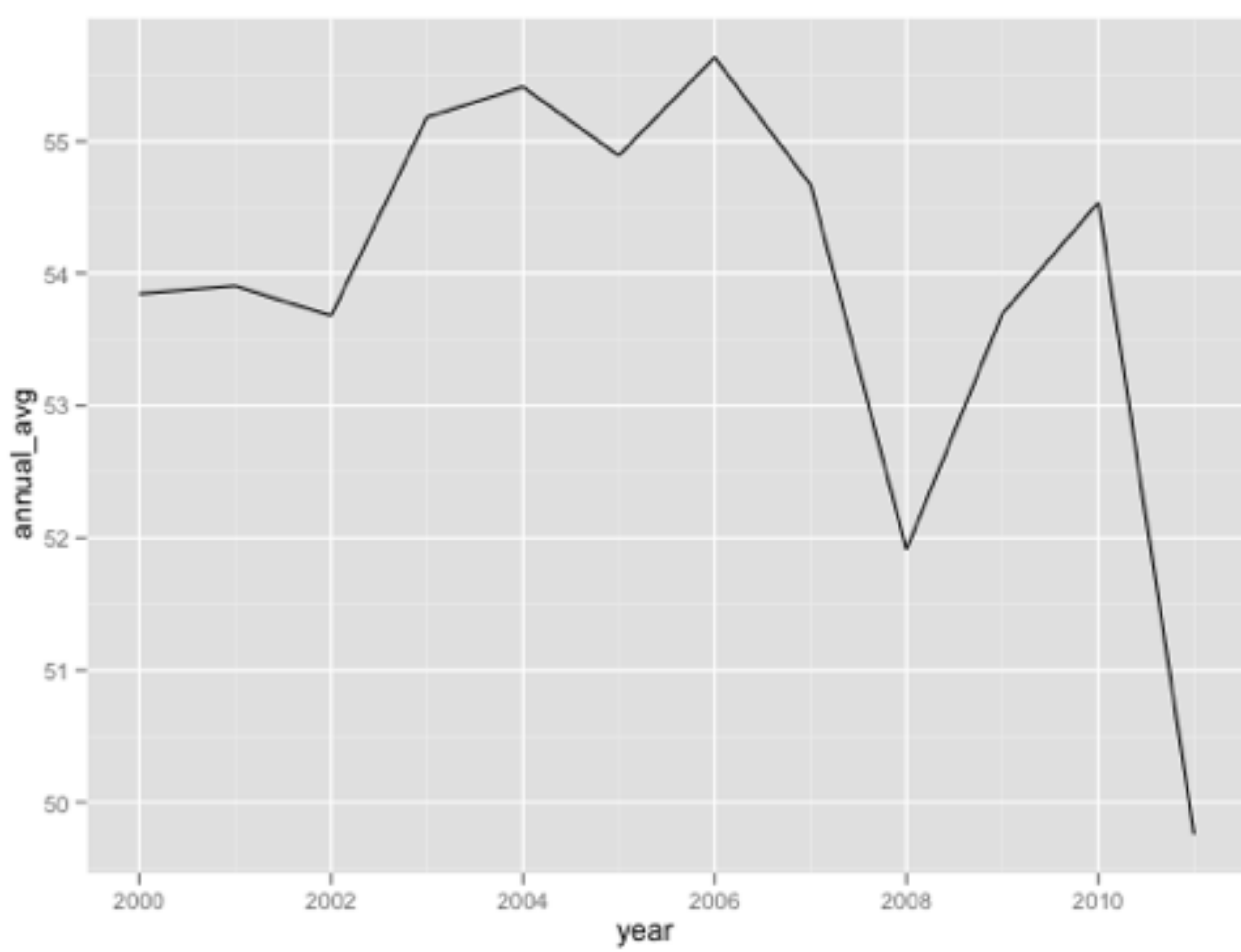
Missing values are given
NA in fitted and residuals

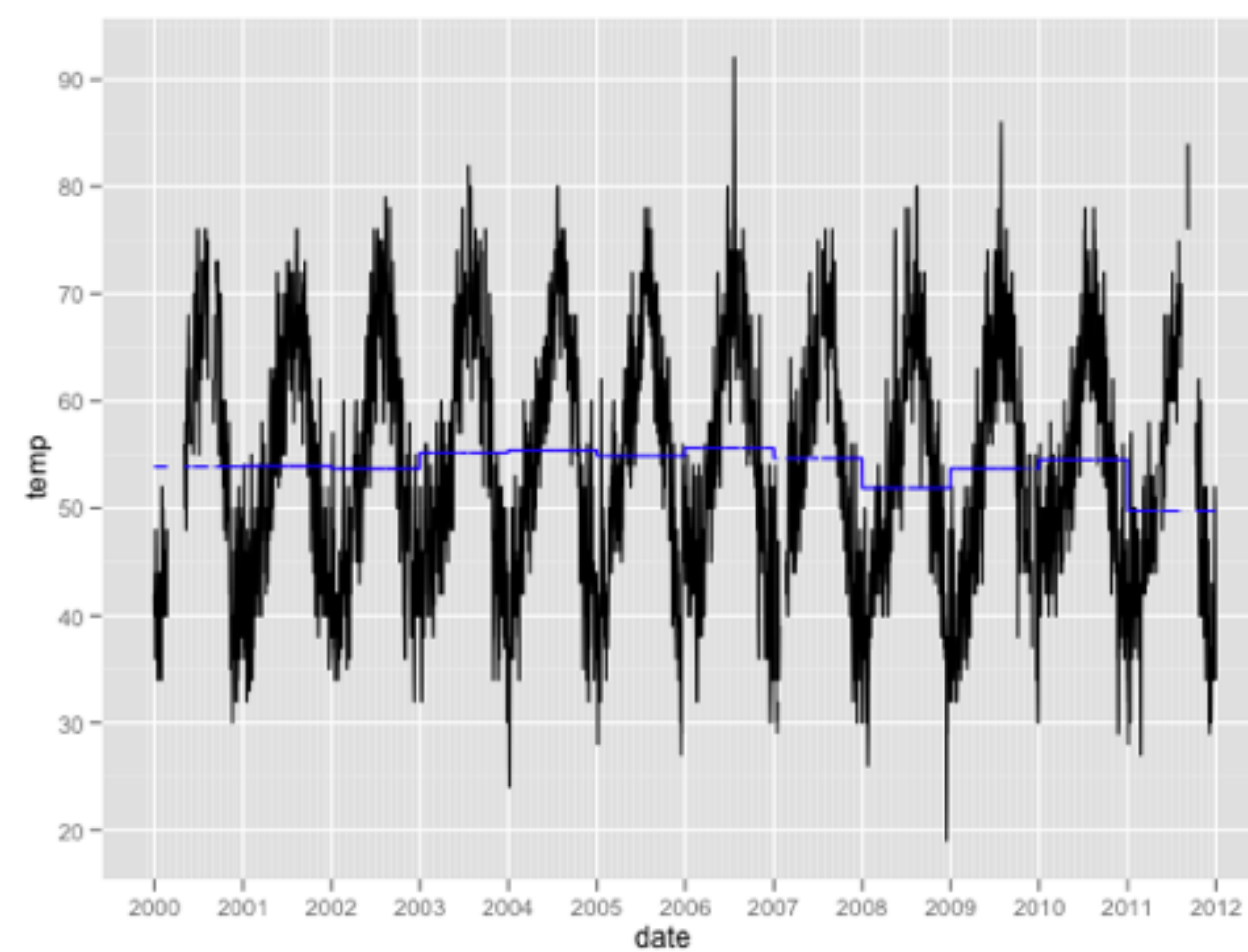
Aggregate

```
corv$annual_avg <- fitted(year_fit)
```

```
qplot(year, annual_avg, data = corv,
  geom = "line")
```

```
qplot(date, temp, data = corv,
    geom = "line") +
  geom_line(aes(y = annual_avg),
    colour = "blue")
```





Your turn

Discuss with your neighbor:

What are the dangers of aggregating over a fine time resolution?

Outliers - average sensitive to single or a few observations (median, maybe keep other summaries)

Lumped all variation into “noise”, ignoring variation below the level of aggregation.

Overlook short term changes

If the true “seasonality” isn’t at the same frequency as the aggregation level, weird artifacts.

Missing values?

How many observations went into summary?

Smooth

Moving average `filter`

Linear regression `lm`

Local regression `loess`

many more … (`gam`)

Smooth: Moving average `filter`

A moving average takes the average of n consecutive values.

If it is centered, take the average of 2q+1 consecutive values giving a half weight to numbers on each end.

For daily data, I might do a moving average over 7 days with equal weights (this is not a centered moving average).

Or, q = 15, approximately a monthly moving average

q =?, approximately a yearly moving average

Smooth: Moving average `filter`

`filter` works on `ts` objects, so let's get our temperatures in to `ts` properly.

There were dates missing from our series:

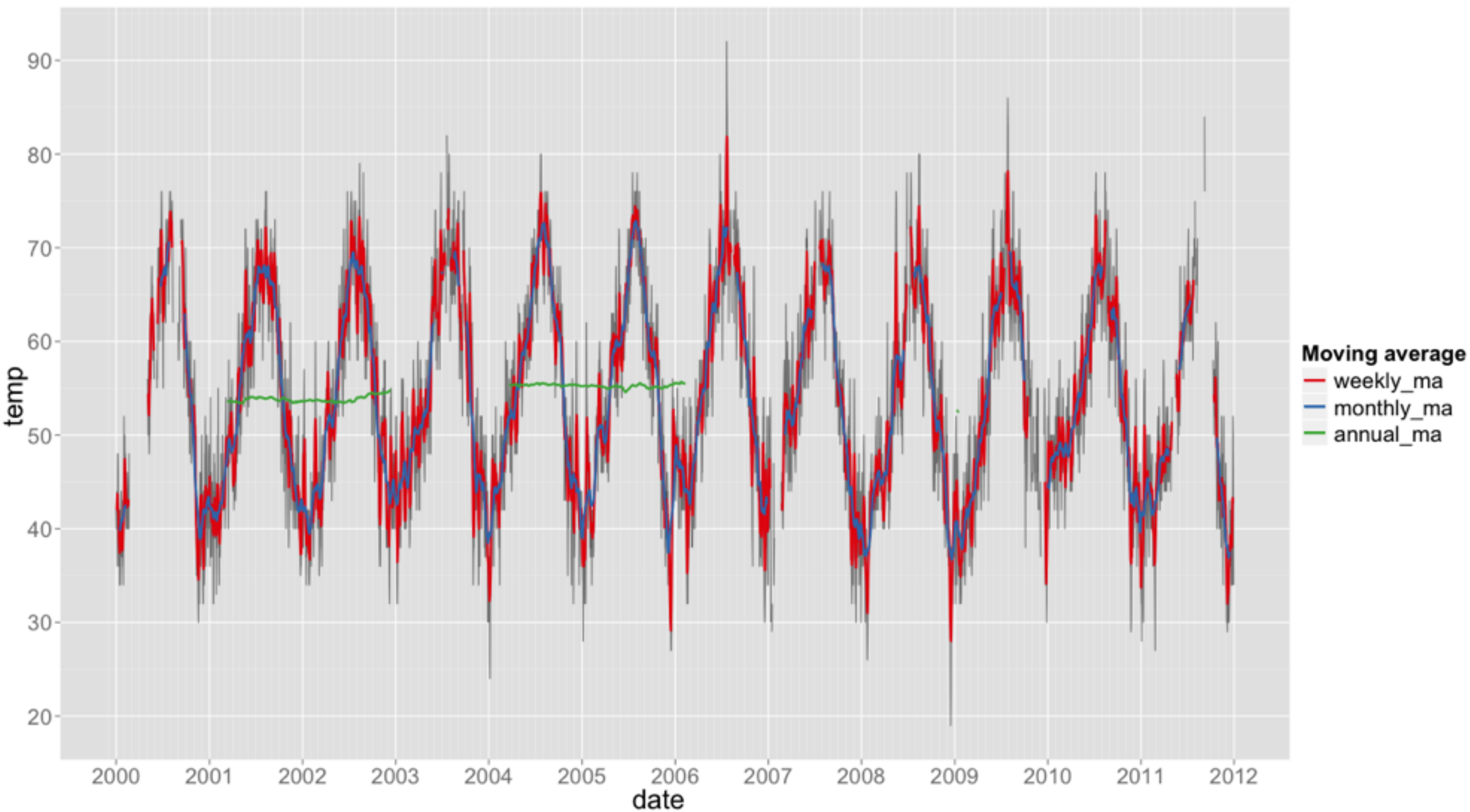
```
all_days <- data.frame(date = seq(ymd("2000/01/01"),
                  ymd("2013/12/31"),  by = "day"))
# merge them together
corv <- inner_join(corv, all_days)
# force it to be in time order
corv <- corv[order(corv$date), ]
# make a ts object
corv_ts <- ts(corv$temp, start = 2000, freq = 365.25)
```

Smooth: Moving average `filter`

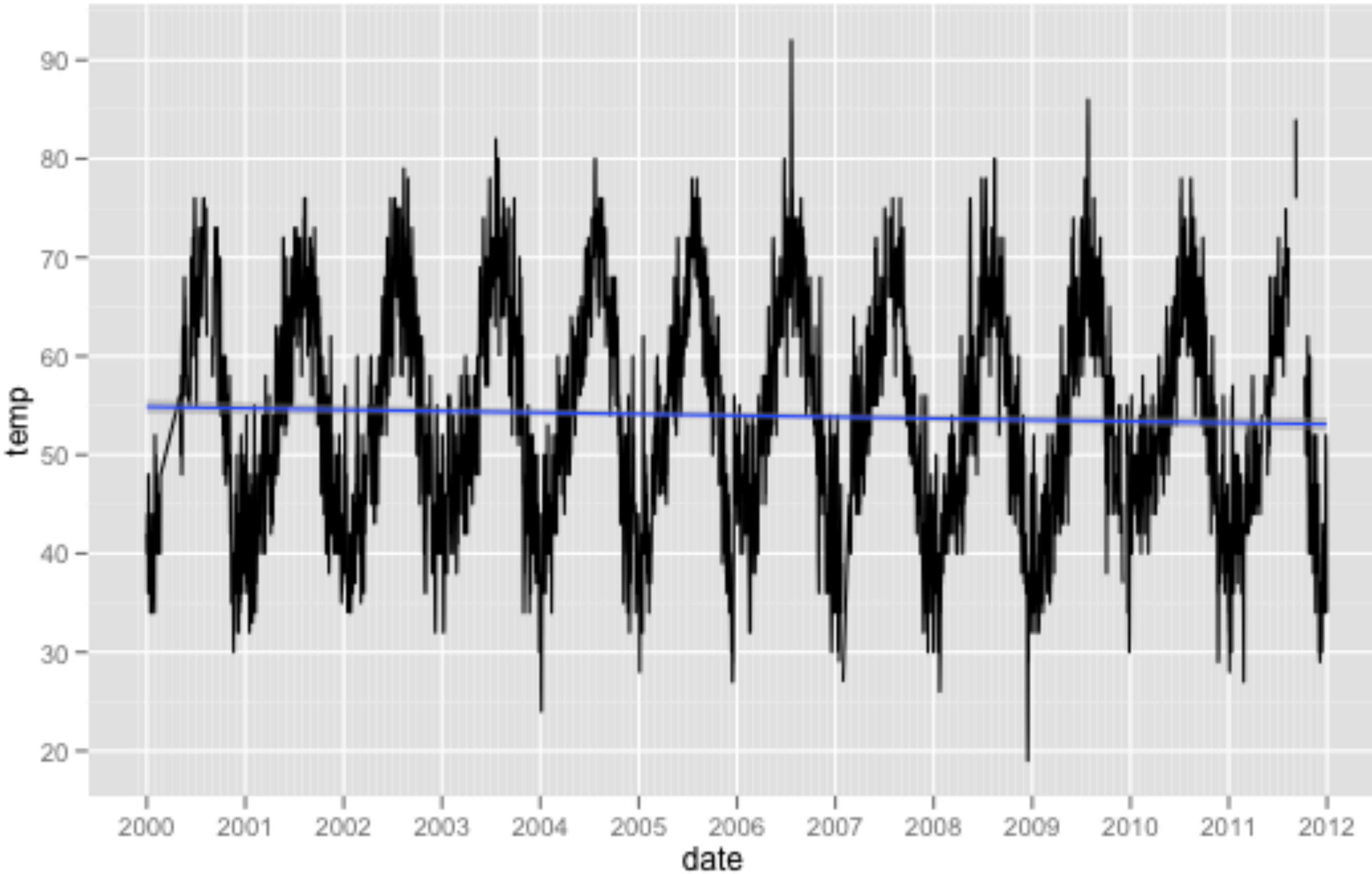
```
n <- 7 # we have daily data so this is a week
corv$weekly_ma <- filter(corv_ts, filter = rep(1, n)/n)
n <- 30 # approximately a month
corv$monthly_ma <- filter(corv_ts, filter = c(1/2, rep(1, n-1), 1/2)/n)
n <- 365 # approximately a year
corv$annual_ma <- filter(corv_ts, filter = rep(1, n)/n)

qplot(date, temp, data = corv, geom = "line", alpha = I(0.5)) +
  geom_line(aes(y = weekly_ma, colour = "weekly_ma"), size = 1) +
  geom_line(aes(y = monthly_ma, colour = "monthly_ma"), size = 1) +
  geom_line(aes(y = annual_ma, colour = "annual_ma"), size = 1) +
  scale_colour_brewer("Moving average", pal = "Set1") +
  big_font
```

Smooth: Moving average `filter`

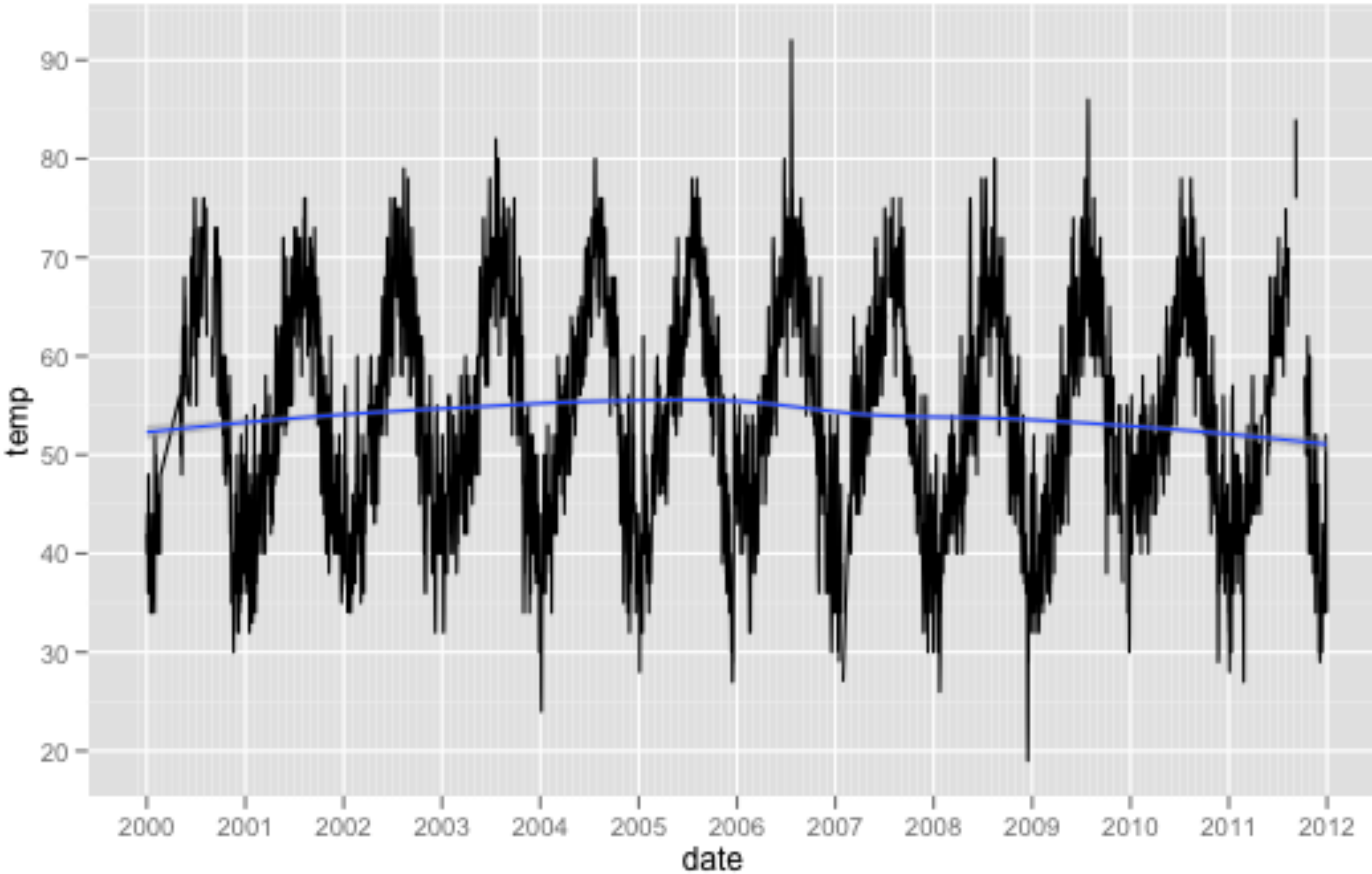


Smooth: Linear regression `lm`



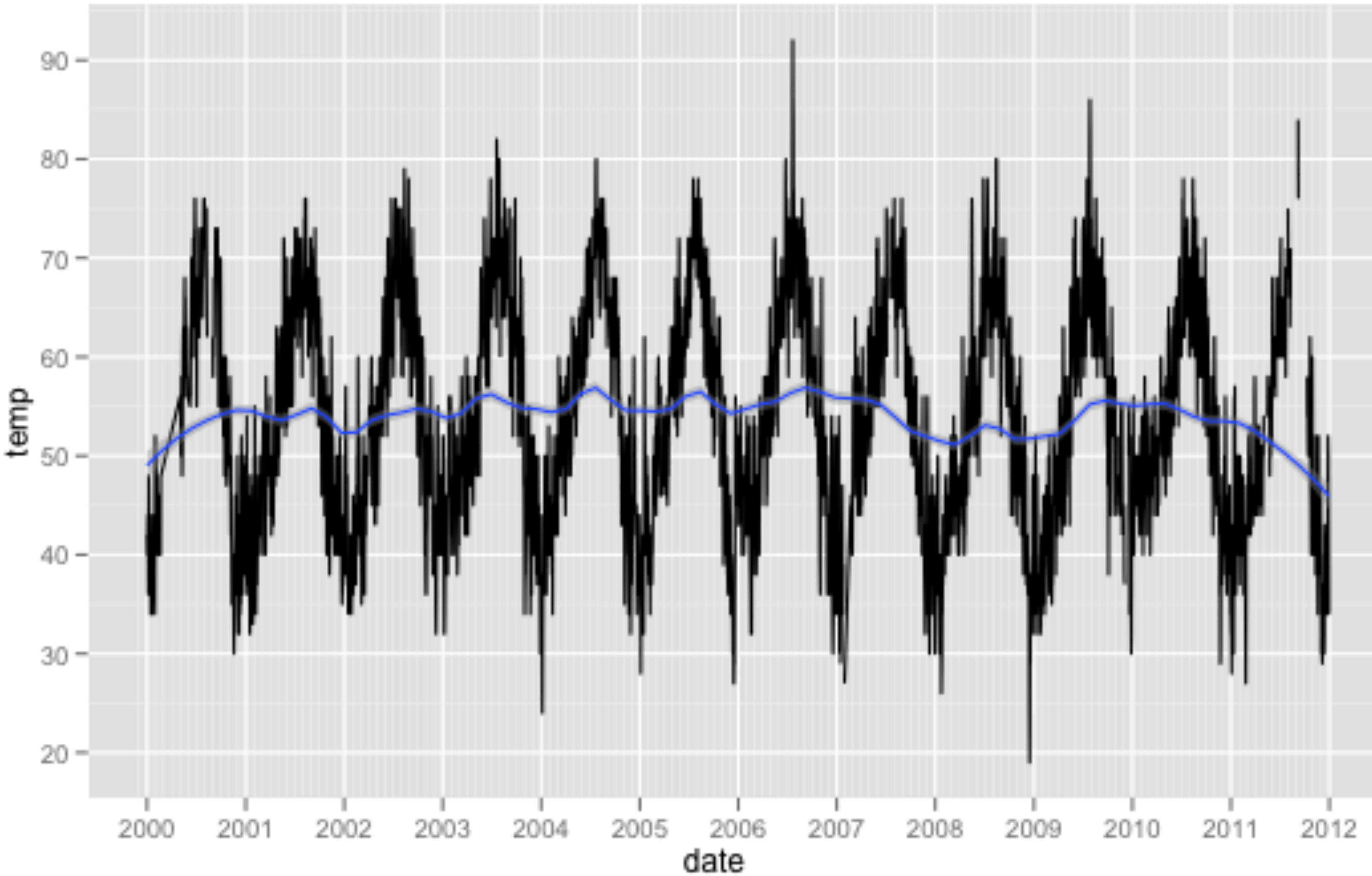

```
qplot(date, temp, data = corv, geom = "line") +
  geom_smooth(method = "lm", se = FALSE)
```

Smooth: Local regression loess




```
qplot(date, temp, data = corv, geom = "line") +
  geom_smooth(method = "loess")
```

Smooth: Local regression `loess`




```
qplot(date, temp, data = corv, geom = "line") +
  geom_smooth(method = "loess", span = 0.2)
```

Your turn

Discuss with your neighbor:

How do you choose a smooth (between methods, or smoothing amount)?

It depends!

Depends on the data, on your goal, on your interest, your purpose for the smooth.

Goal: prediction/forecasting, might choose method based on forecast horizon.

Does ease of interpretation matter?

Rely on defaults, (dangerous), guess and check, trial and error.

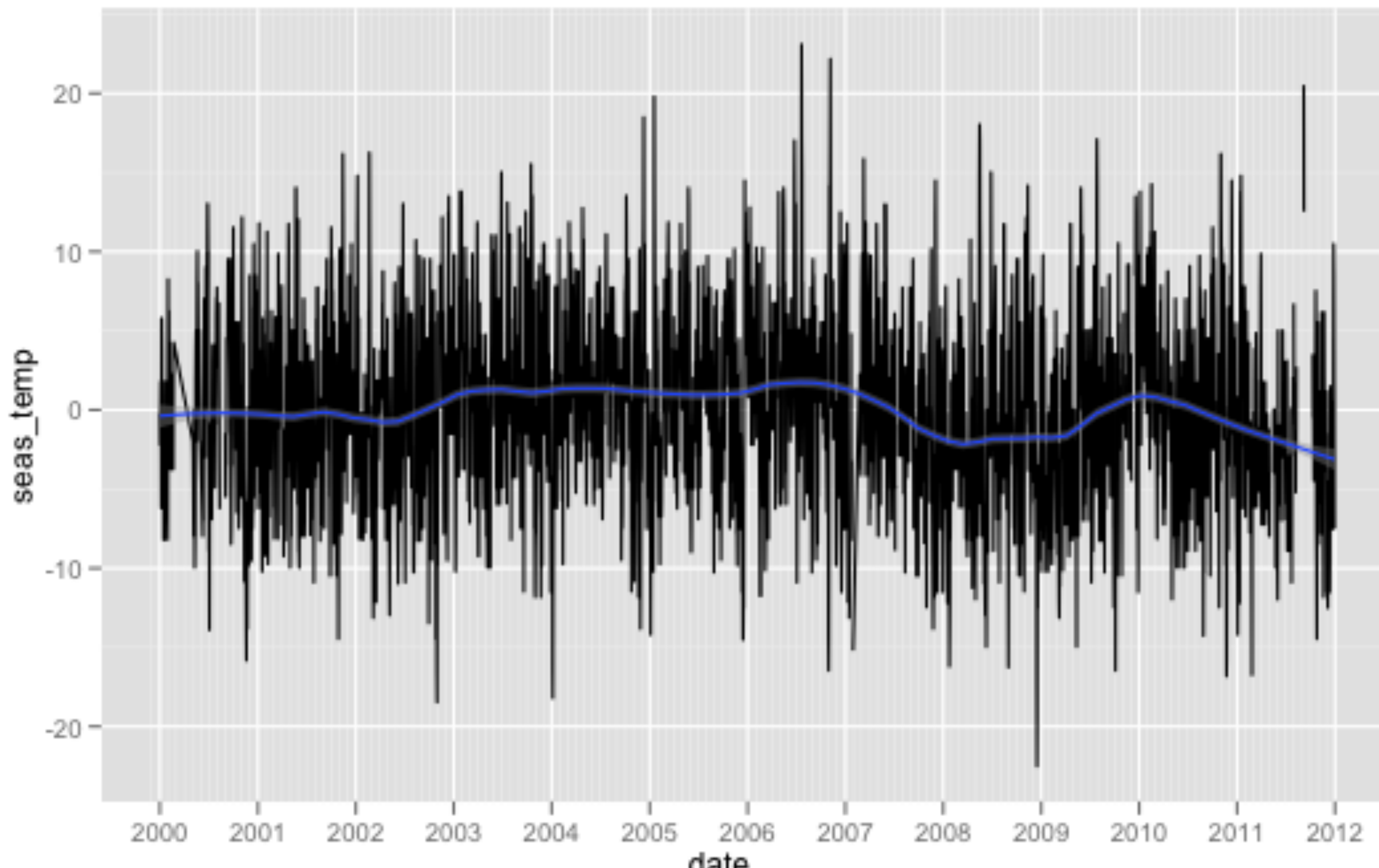
Be wary of “optimal” defaults

Try until it “looks good”

Subject matter knowledge

Subtract

```
month_fit <- lm(temp ~ factor(month), data = corv, na.action = na.exclude)
corv$month_avg <- predict(month_fit)
corv$res <- residuals(month_fit)
```




```
qplot(date, month_avg, data = corv, geom = "line", group = year)
qplot(date, res, data = corv, geom = "line") +
  geom_smooth(method = "loess", span = 0.2)
```

A common decomposition

$$x_t = m_t + s_t + z_t$$

Variable measured at time t | Trend | Seasonality | Noise

$$x_t = m_t s_t z_t$$

$$x_t = m_t s_t + z_t$$

some multiplicative analogs

Suggests a general approach

Describe and model the most obvious part

Subtract it from the series and repeat

When there is just "noise" left, examine it's variance and correlation

Hints for subtracting

Find a model for the part you want to subtract using: `lm`, `loess`, ...

Fit the model (with `na.action = na.exclude`)

Use `predict` to get the model prediction

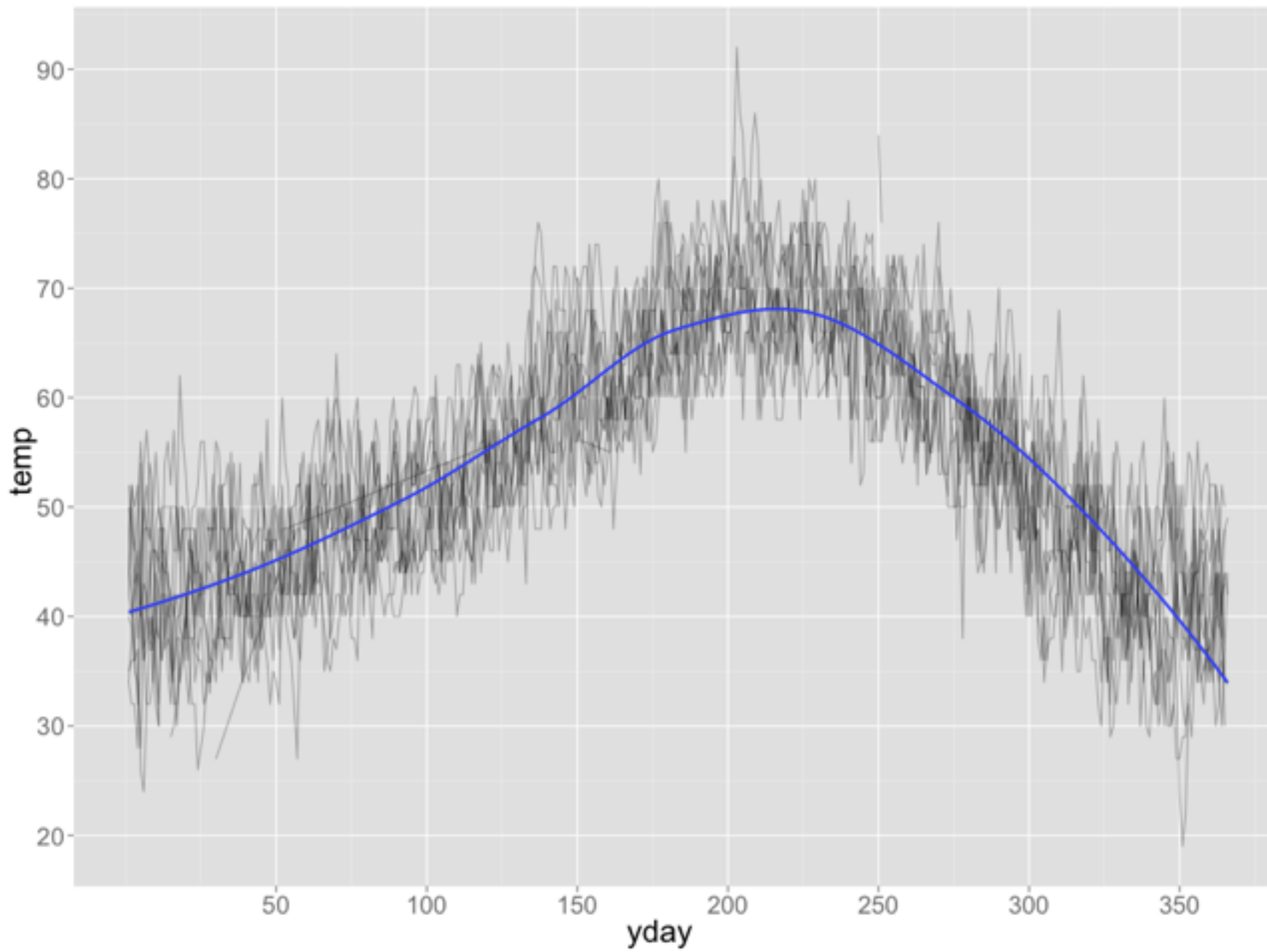
and `residuals` to do the subtraction

If you want to use a moving average, the output from `filter` is the prediction, do the subtraction "by hand"

An example

Corvallis data, the largest variation is from the seasonal pattern.

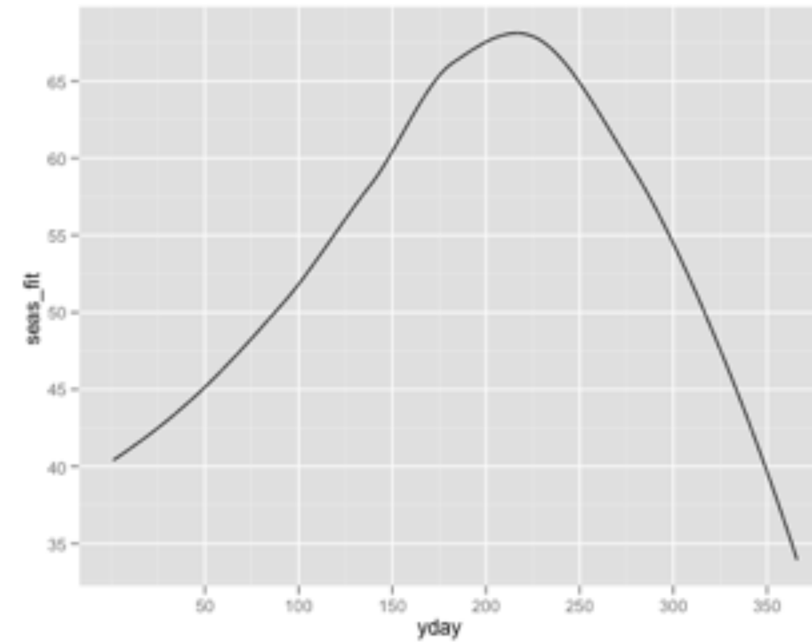
Might be nice to use a smooth.



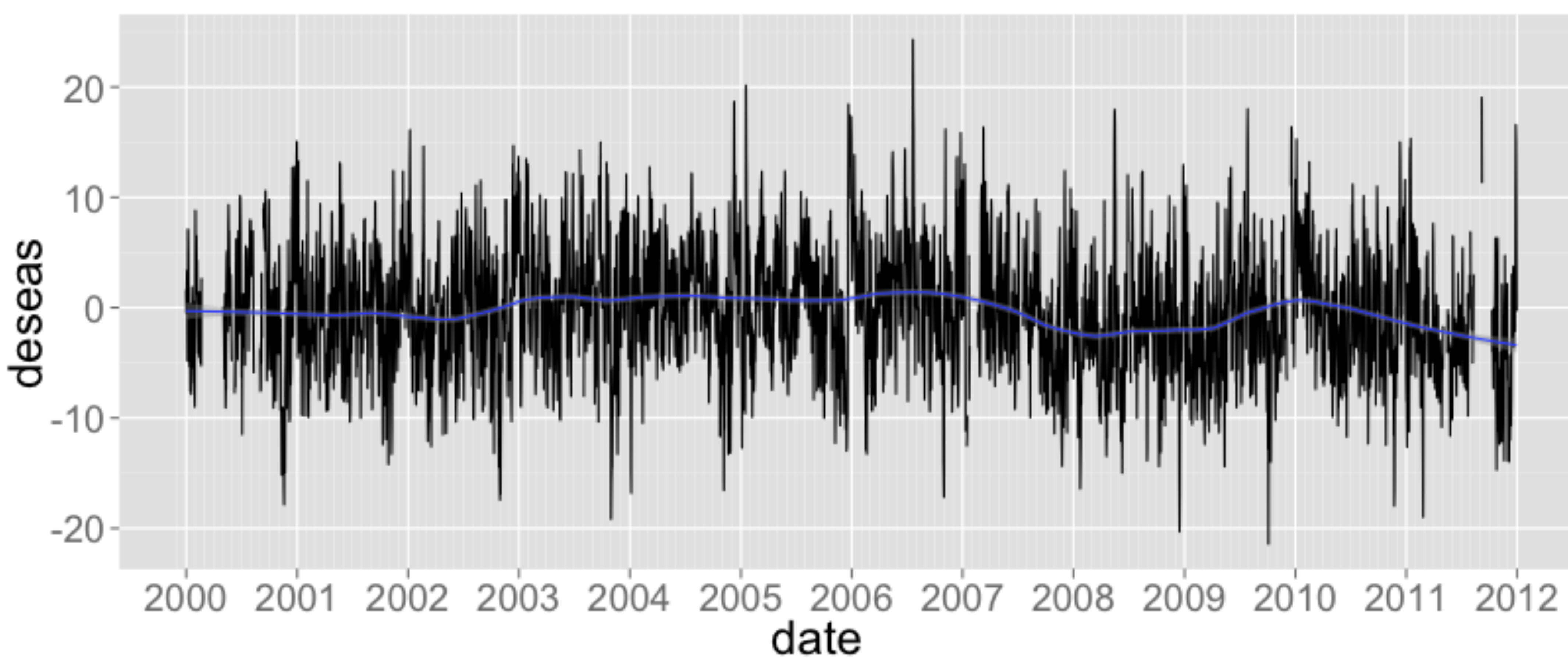

```
qplot(yday, temp, data = corv, geom = "line", group = year, alpha = I(.3)) +
  geom_smooth(method = "loess", aes(group = 1), size = 1)
```

```
seas_fit <- loess(temp ~ yday, data = corv,
                  na.action = na.exclude)
corv$seas_fit <- predict(seas_fit,
                  newdata = corv)
corv$deseas <- residuals(seas_fit)
```

```
qplot(yday, seas_fit, data = corv,
    geom = "line")
```




```
qplot(date, deseas, data = corv, geom = "line") +
  geom_smooth(method = "loess", span = 0.2)
```



Your turn

Where would you start with this series?

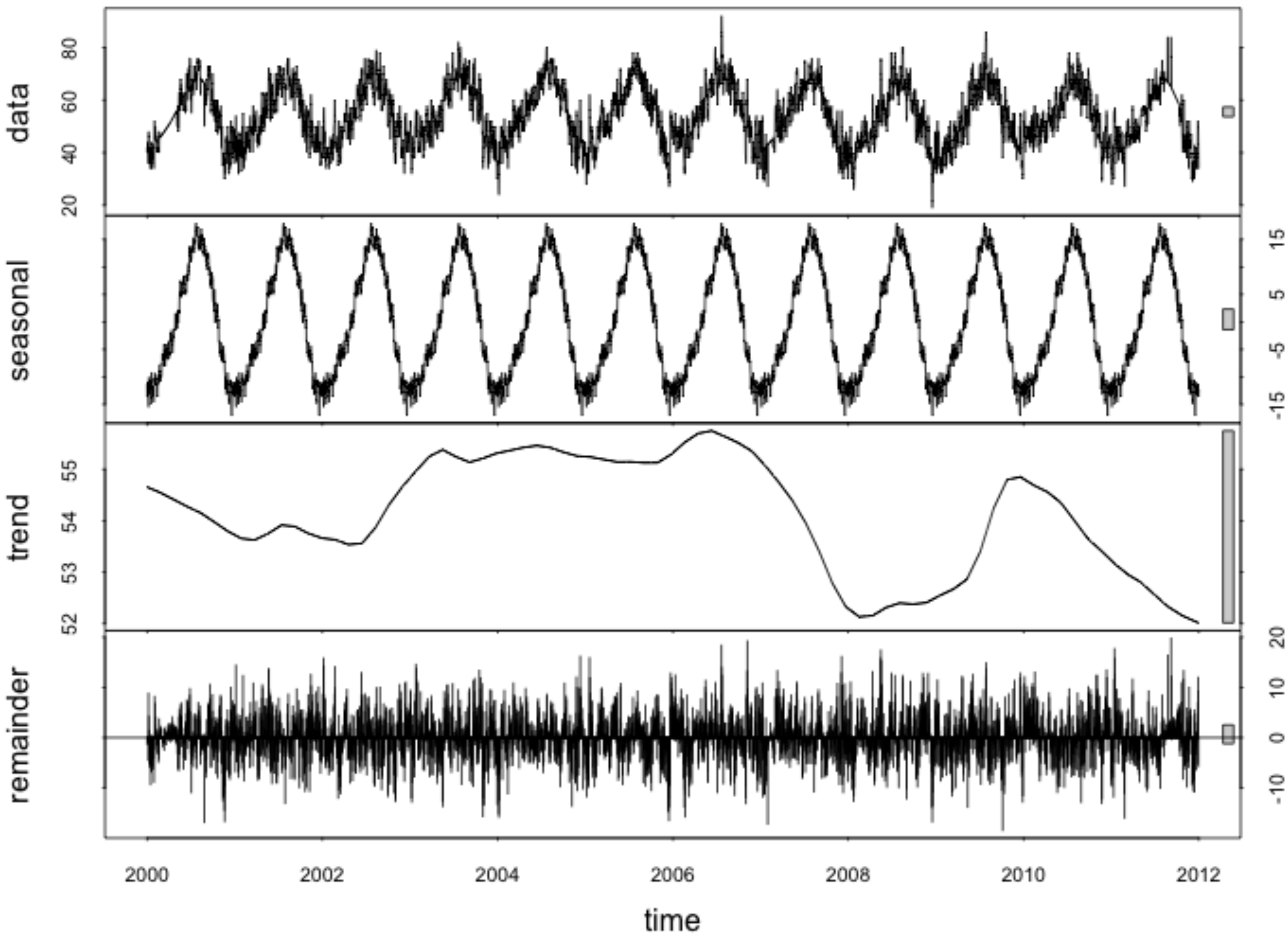
Two automatic approaches

`stl` iteration of local regression smooths

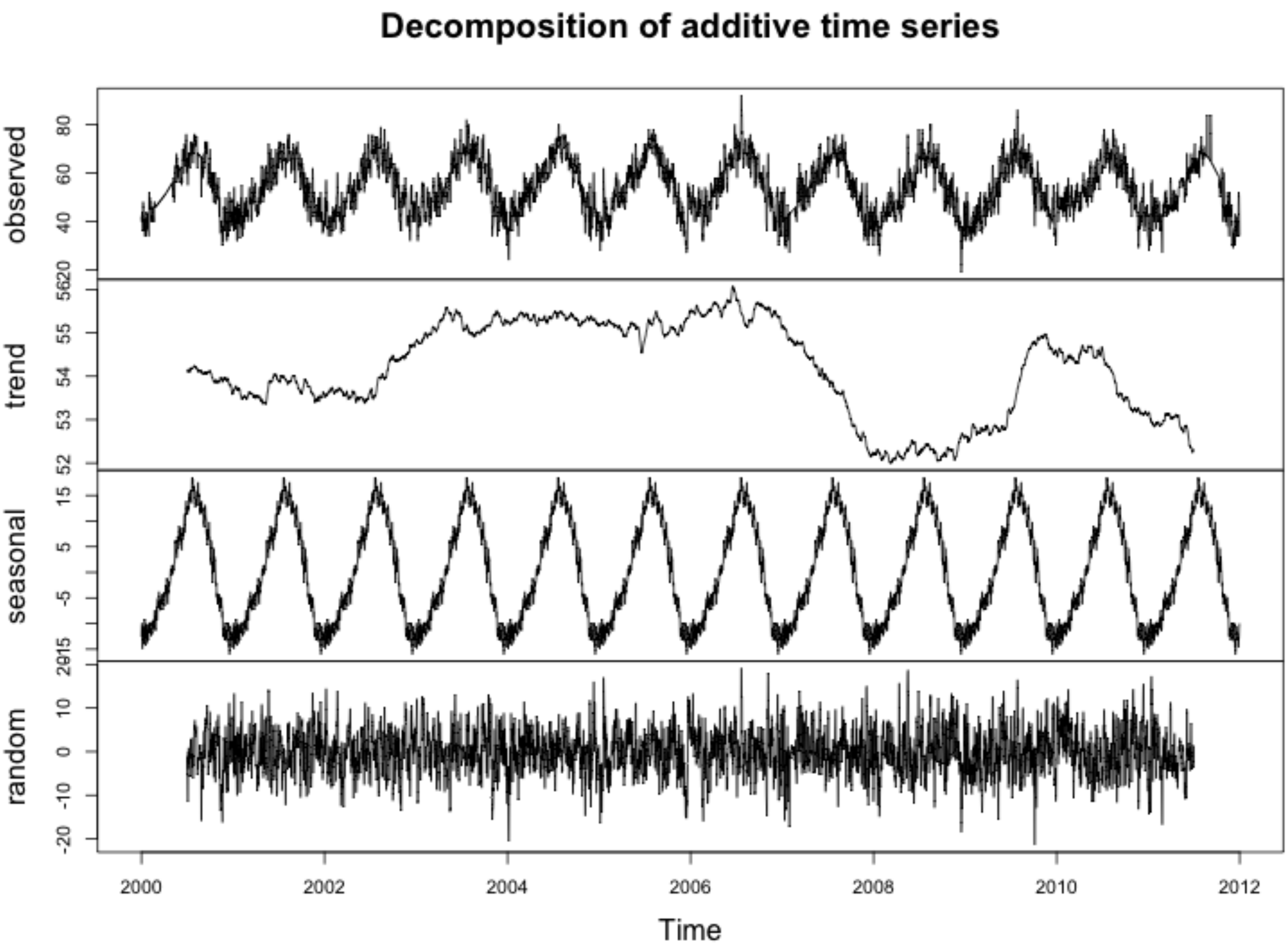
`decompose` iteration of moving averages

neither like missing values

```
plot(stl(corv_ts2, 365.25))
```



```
plot(decompose(corv_ts2))
```



Next time…

Examining variance and correlation