Nokkur grundvallaratriði

Helgi Tómasson
helgito@hi.is

15. janúar 2024

- Kunna AR(1) vel:

$$Y_t = \phi Y_1 + \varepsilon_t$$
$$\text{Dreg } Y_{t-1} \text{ frá báðu megin}$$
$$\Delta Y_t = Y_t - Y_{t-1} = \phi Y_t - Y_{t-1} + \varepsilon_t = (\phi - 1) Y_{t-1} + \varepsilon_t$$

- Berið saman við línulega diffurjöfnu af einföldustu gerð:

$$Dy(t) = y'(t) = ay(t)$$
$$\text{lausnin er: } y(t) = \exp(at)y(0),$$
$$\text{Ef } a < 0 \text{ stefnir lausnin á 0, fasta ef } a = 0 \text{ og veldisvöxt ef } a > 0$$

- AR(1) líkan er strjál-tíma útgáfa af fyrstu gráðu diffurjöfnu.
- Það er verið að lýsa hreyfimynstri (*dynamics*)

Tveir ferlar (processar) að skilja vel

- White-noise, ε_t (AR(1), með $\phi = 0$),

$$E(\varepsilon_t) = 0, \quad V(\varepsilon_t) = \sigma^2, \gamma(k) = E(\varepsilon_t \varepsilon_{t-k} = 0.$$

- Random-walk, (AR(1), með $\phi = 1$, unit-root),

$$Y_t = Y_{t-1} + \varepsilon_t = \sum \varepsilon_{t-k}, \quad \text{summa af liðnum noise}$$

$$\text{Ef } Y(0) = 0 \text{ þá er}$$

$$E(Y(t)|Y(0) = 0) = 0, \quad V(Y(t)|Y(0) = 0),$$

$$\gamma_k = E(Y_t Y_{t-k}|Y(0) = 0) = \underbrace{\sigma^2(t-k)}_{\text{Fall af } t}$$

- Random-walk er því ekki stationary, varíans stefnir á óendanlegt.
- Þegar mælingar á tímaröð eru skoðaðar þarf að hugleiða, líkjast mælingar, white-noise, random-walk er virðist vera einhvers konar hreyfimynstur þannig að kerfið virðist leita í jafnvægi.

Upprifjun

- Hver er munurinn á $\bar{X}$, $\bar{x}$ og μ?
- Allt er þetta kallað meðaltal. $\bar{X}$ og $\bar{x}$ eru úrtaksmeðaltal. $\bar{X}$ er random-breyta (hending) og hefur einhverja dreifingu. $\bar{x}$ er gildi sem hendingin $\bar{X}$ tekur, þ.e. reiknað meðaltal á mælingunum $(x_1, \ldots, x_n)$. Þetta er því einhver tala. Sanna meðaltalið (oftast óþekkt) líka einhver tala μ.
- Bæði $\bar{X}$ og $\bar{x}$ eru táknuð með $\hat{\mu}$. Túlkunin á $\hat{\mu}$ ákvarðast af samhenginu.
- Ef $\boldsymbol{X}$ er vektor-hending með væntanlegt gildi, $(\boldsymbol{X}) = \boldsymbol{\mu}$ og covaríans-fylki $V(\boldsymbol{X}) = \boldsymbol{\Sigma}$ þá gildir að:

$$E(A\boldsymbol{X}) = A\boldsymbol{\mu}, \quad V(A\boldsymbol{X}) = A\boldsymbol{\Sigma}A'$$

Spurious regression

- Það þarf að gæta varúðar þegar „besta-línan "í tímaröðum
- Sjá frægt aðvörunardæmi https://helgito.hi.is/spurious-yule.pdf
- Stundum er hægt að leiðrétta námkvæmni í mati með reglu Bartlett

Hugleiðingar um AR(1)

Í dæmum 1-3 er gengið út frá:

$$X_t = \phi X_{t-1} + \varepsilon_t,$$

þar sem ε_t er $WN(0, \sigma^2)$ og mælingar eru, $X_1, \ldots, X_T$.

1. Sýnið að

$$\hat{\phi}_{LS} = \frac{\sum_{t=2}^{T} X_t X_{t-1}}{\sum_{t=2}^{T} X_{t-1}^2}.$$

2. Á hvað dreifingu stefnir dreifing $\hat{\phi}_{LS}$ ef $|\phi| < 1$.
3. Hvað getum við sagt um dreifingu $\hat{\phi}_{LS}$ ef $\phi = 1$?

Svörin við þessu veita innsæi í „unit-root prófin sem þið hafið lært um í hagrannsóknum. Til að geta leyst þessi dæmi er gott að kunna svolítið um samleitni hugtök svo sem samleitni í dreifingu og plim.

4. Rifjið upp stokastíska regressora í regressionslíkönum.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad E(\varepsilon) = 0, \quad V(\boldsymbol{\varepsilon}) = \sigma^2 I.$$

   Ef ákveðnar forsendur gilda þá $\hat{\boldsymbol{\beta}} \xrightarrow{d} N(\boldsymbol{\beta}, \sigma^2 Q)$. Hvaða fyki er Q? Hvað brestur í spurious regression dæminu þegar:

$$y_t = \beta x_t \varepsilon_t,$$

   þar sem x_t er random walk og ε_t er white-noise óháð x_t.

5. Hermið MA(1), t.d. í spreadsheet eða R. Gerið ráð fyrir að vitað sé að $\sigma = 1$. Metið θ með formúlunni $\hat{\theta} = \frac{1}{n}\sum_{t=2}^{T} x_t x_{t-1}$.
   Hvað heitir þessi matsaðferð?. Getum við útvíkkað hana fyrir t.d. AR(p)? Lesið um Yule-Walker jöfnur á netinu.

Um sjálffylgni stationary ferla

- Sjálfylni fall AR(1) er ϕ^k, þ.e. veldisdempun.
- Sjálffylgnifall MA(1) er $\frac{\theta^2}{1+\theta^2}$ fyrir k=1 og 0 ef k>1.
- Almennt gildir að sjálffylgnifall AR líkana er veldisdempað en sjálffylgnifall MA fer snögglega niður í 0.
- Hegðun PACF, $E(Y_t Y_{t-k} | Y_{t-1}, \ldots, Y_{t-k+1})$ er öfug, þ.e. PACF fer snögg niður í 0 fyrir AR líköng en er veldisdempað.
- Sjálffylgni ARMA er veldisdempuð frá ákveðnu k.

Sundurliðu tímaraðar

- Í kennslubókum er sagt:

$$y_t = \text{trend} + \text{season} + \text{cycle} + \text{irregular}$$

- Trend, langtímaþróun getur verið ýmist random, eins og t.d random-walk, eða deterministic, þ.e. útreiknanlegt fall af tíma, eða blanda af þessu tvennu. T.d.;

$$Y_t = \alpha + \beta t + Y_{t-1} + \varepsilon_t$$

- Á sama hátt getur season,verið deterministísk,stókastísk eða blanda.
- Sama á við um cycle verið blanda af t.d. þríhyrngaföllum af t og ARMA-liðum.
- Irregular hlutinn á að vera óspáanlegi hluti ferlisins.

Box-Jenkins nálgun

- Tímaraðagögn virðast oft ekki vera mælingar á stationary ferlum. Það þarf að finna einhvern stationary þátt í ferlinu. BJ stungu upp á að hreinsa burt deterministíska þætti og gera ARIMA líkan fyrir random þáttinn.
- ARIMA(p,d,q,P,D,Q,season=S) var hugsað þannig:

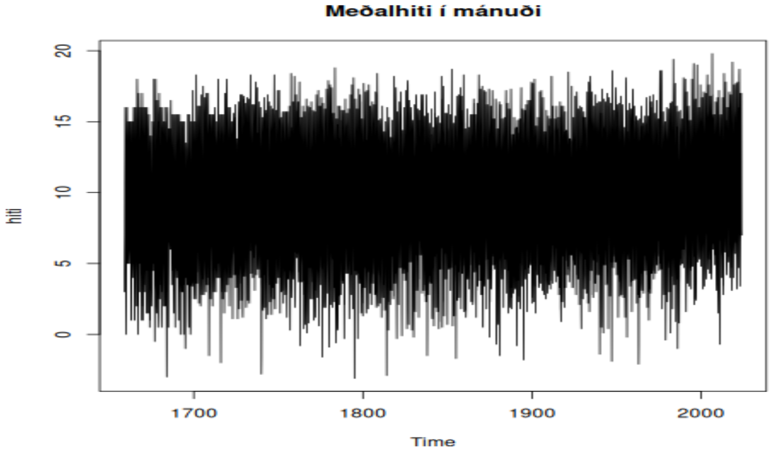
$$
\begin{aligned}
&\Delta^d \Delta_S^D Y_t, \text{ er ARMA}\\
&\Delta = 1 - L, \quad \Delta_S = (1 - L^S),\\
&p = \text{fjöldi AR tafa}, \quad q = \text{fjöldi MA tafa}\\
&P = \text{fjöldi AR seasonal tafa}, \quad Q = \text{fjöldi MA seasonal tafa}\\
&S = \text{lengd árstíðar, 12 ef gögn eru mánaðarleg}
\end{aligned}
$$

1. Veljð heppilegt, variance-stabilizing transform. Í haggögnum er þetta oftas logaritmi. Plottið mynd af röð, giskið á trendstúktúr. Giskið á d, D, p, q, P, Q. Við gisk á d og D stungu BJ upp á að lágmarka $V(\widehat{\Delta^d \Delta^D} y_t) =$(þið hafið lært um unit-root próf sem eru öðruvísi. Giskið á p, q, P, Q byggist á því að skoða úrtaks-ACF, úrtaks-PACF. Kannski CUMSUM og CUMSUMSQ á frávikum frá meðaltali.
2. Metið parametra með einverri aðferð.
3. Skoðið frávíkin, ACF, PACF, CUMSUM, CUMSUMSQ ofl. (kannski spektral eiginleika,ekki búinn að fara í).Ef vel hefur tekist til eiga frávikin að líkjast white-noie.
4. Reiknið spá og áætlið stærð spáskekkju.

Hugleiðið hvernig þróast spáskekkja fyrir random-walk? Principle parsimony (nískuprinsípið), hafið p,q,P,Q,d,D lágar tölur.

Sýnidæmi

Hiti í Englandi 1659 til 2023

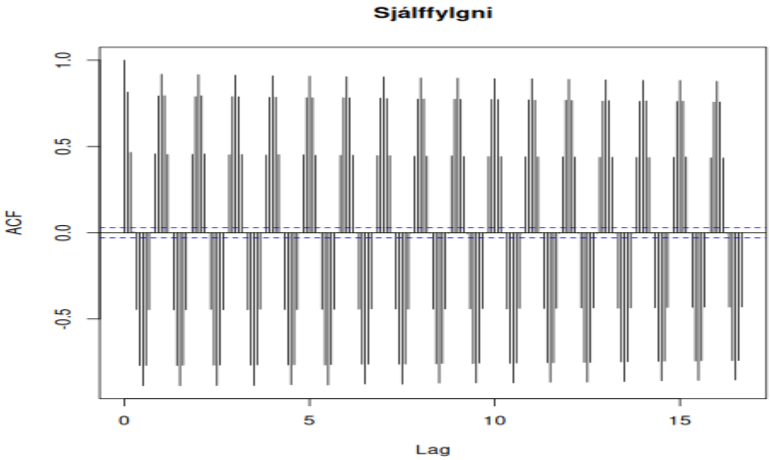


$\widehat{V(y_t)} = 22.9,\ \widehat{V(\Delta y_t)} = 8.5,\ \widehat{V(\Delta_{12} y_t)} = 3.6,\ \widehat{V(\Delta\Delta_{12} y_t)} = 5.4$

Gögnin eru sótt af:

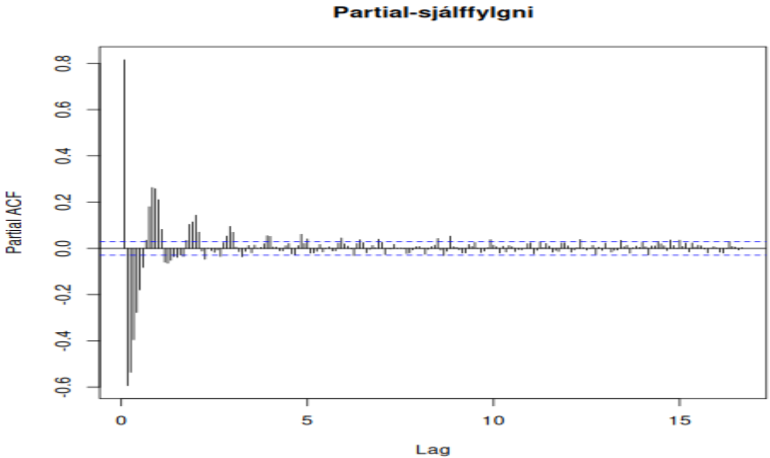
https://www.metoffice.gov.uk/hadobs/hadcet/data/download.html

ACF



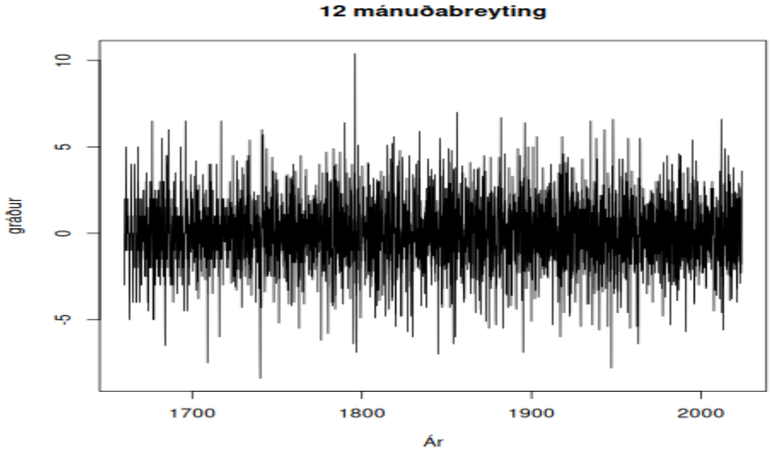
Athyglivert að sjálffylgni á laggi 12 virðist minnka mjög hægt. .9187743 0.9164271 0.9131262 0.9094351 0.9076567 0.9043007, meira um það síðar.

PACF



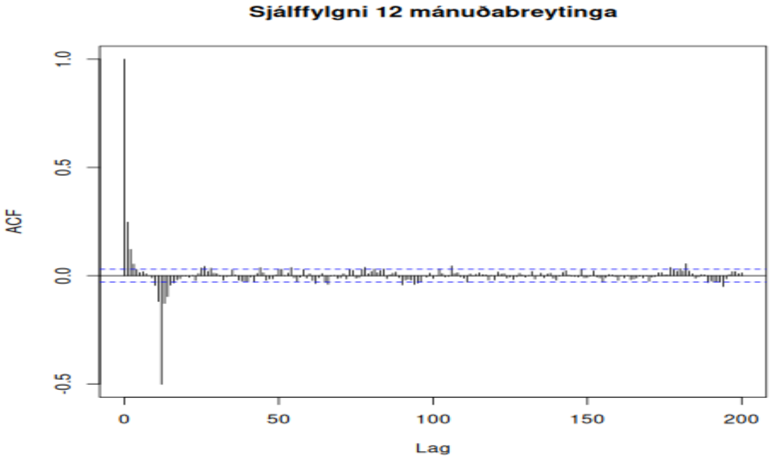
Athyglivert að saga sem er eldri en ca. 5 ára virðist ekki skipta máli.

Skoða diffraða röð, $\Delta_{12}y_t$, 12 mánaða breyting, frá 1660 til 2023



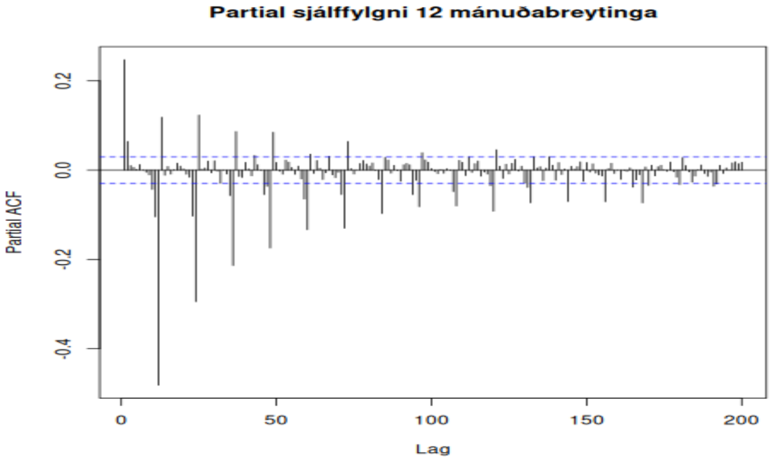
$\bar{y}$=0.006, staðalfrávik=1.9. t-gildi=0.21, Er þetta marktækt frábrugðið 0? (Marktækni varasamt hugtak).æ

ACF



Sjálffylgni 12 mánuðabreytinga
ACF
1.0
0.5
0.0
-0.5
0
50
100
150
200
Lag

PACF



Þessar tvær myndir gefa í skyn að MA framsetning sé heppileg.

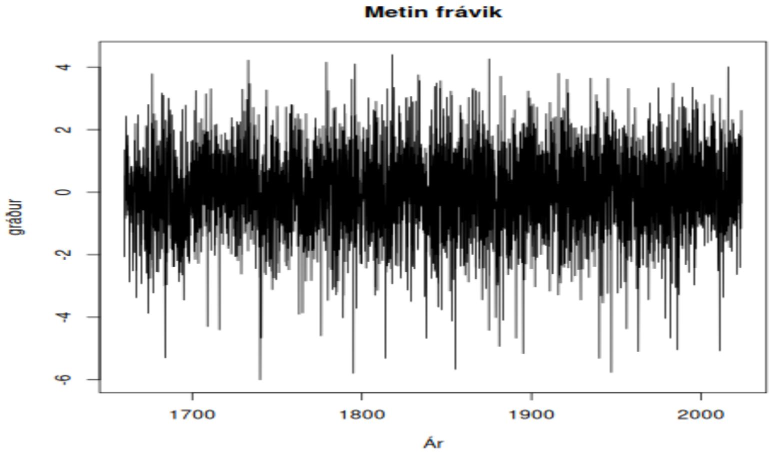
MA líkan

Sting upp á:

$$\Delta_{12} Y_t = 0.0042 + 0.2627\varepsilon_{t-1} + 0.1107\varepsilon_{t-2} + 0.0092\varepsilon_{t-3}$$
$$+0.0164\varepsilon_{t-11} - 0.9833\varepsilon_{t-12} - 0.2335\varepsilon_{t-13} - 0.0953\varepsilon_{t-14} + \varepsilon_t$$

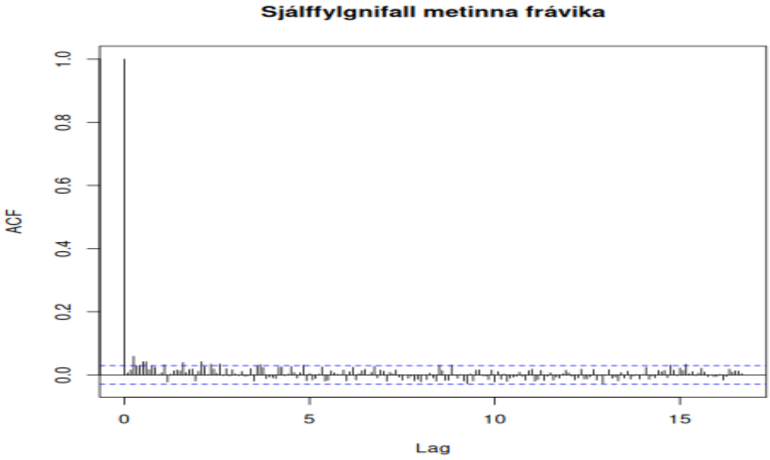
$\bar{\hat{\varepsilon}}_t = -0.005, \quad \widehat{V(\hat{\varepsilon}_t)} = 1.72$ Óútskýrðurbreytileiki er því 1.7/3.6.
Þ.e. helmingur að breytileika í 12-mánuða breytingum skýrist með ARMA sveiflum.

Metin frávik (spáskekkjur)

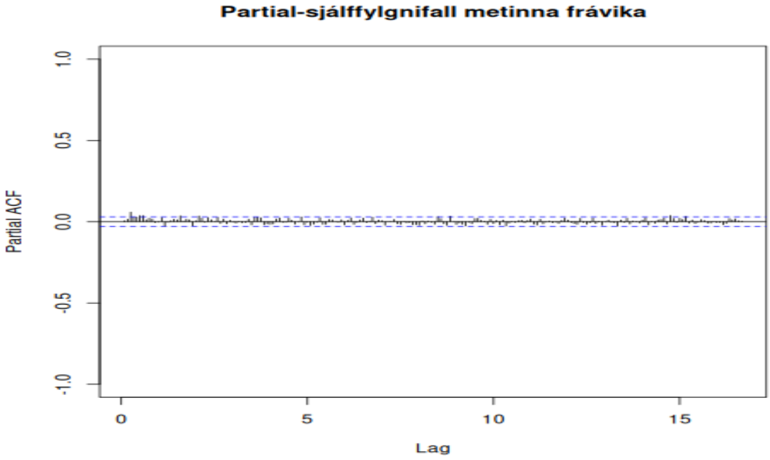


Breyta sem vantar í líkanið þarf að geta skýrt þessar hreyfingar.

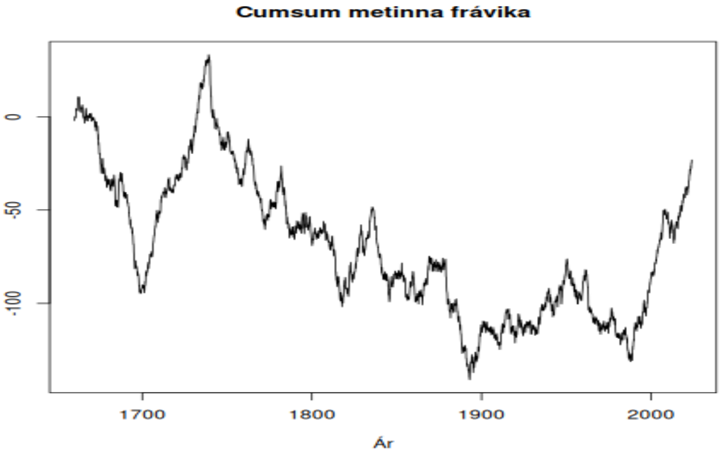
ACF metinna frávika



PACF metinna frávika

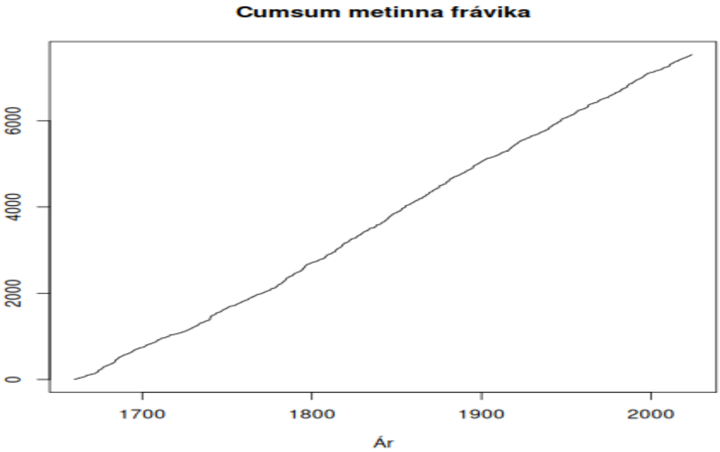


CUMSUM metinna frávika



Þessi ferill á að líta út eins random-walk.

CUMSUMSQ metinna frávika



Ef þessi ferill er bein lína þá er það vísbending um einstdreifni (homoskedacity). Þ.e. breytileiki frávika er fasti.

Umræður:Hvað er rangt við þetta?

1. Líkanið er náttúrulega fáranlega einfalt. Ein diffrun og nokkrir metnir parametrar ná að lýsa miklum hluta hreyfimynsturs,
2. Hagfræðingarnir upp úr 1970 sögðu að með þessari nálgun missti spáin allt samband við levelið. Þ.e. ég fáið notahæfa spá í rúmlega eitt ár og eftir það sé alltaf spáð fastri tölu (mér sýnist hlýnun) og að spáskekkjan mun vaxa. Vöxtur spáskekkjunnar er algerlega órýmileg.
3. Þ.e. að það sé órýmilegt að vera með ARIMA(p,0,q,P,1,Q, season=12). Með þeim rökum náttúrulega enn órýmilegra að vera með ARIMA(p,1,q, P,1,Q,season=12), eins og ég hef séð í vísindagrein.
4. Algerlega rétt hjá hagfræðingunum. Það á að forðast ofdiffrun (hver er varíans á diffruðum white-noise?).
5. Spurious-regression vandinn sem getur komið upp þegar tímaraðastrúktúr er til staðar er til vandræða. Venjuleg tölfræði gildir um stationary tímaraðir þegar leiðrétt hefur verið fyrir sjálffylgni.
6. Error-correction (tengir saman level og breytingar) og co-integration sem hefst upp úr 1980 er því mikið gegnumbrot.

7. Ný og flókin teoría, long-memory líkön (kafli 2.11 í bók) tekur fyrir ferla þar sem sjálffylgnin deyr hægar út en veldisvöxtur segir til um. Hugmynd er að diffrunin, d, eigi ekki endilega að vera heil tala.
8. Í ensku hitaröðinni er árstíðasveiflan áberandi. Það þarf að leiðrétta einhvern veginn. Dagsvik og félagar (2020) gera það (minnir mig) með því að reikna fljótandi meðaltal yfir 12 mánuði í hitagögnum út um allan hnöttinn. Dagsvik og félagar meta síðan long-memory líkan á 100 veðurstöðvum í ca. 200 ár.
9. Almennt finnst mér að fólk þurfi að hafa varann á sér þegar það reiknar fljótandi meðaltöl í tímaröðum. Ég tek það fyrir síðar, þegar við skoðum spektral-teoríuna. Nóbelsverðlaunahafi í hagfræði (Kuznetz) bjó hugsanlega til spurious-cycle með fljótandi meðaltölum Meira um það síðar.
10. Long-memory líkönin eru erfið, bæði stærðfræðilega (ferlar ekki Markov né martingale), og reiknfræðilega. Það er seinlegt að reikna bestu spá og sennileikafall (likelihood).

11. Stefni á að sýna nálgun á ensku gögnunum sem er frábrugðin Dagsvik), önnur árstíðaleiðrétting og annað langtíma líkan. Dagsvik notar fractional-gaussian-noise (fgn), en ég ætla að sýna ARFIMA nálgun (því ég kanna ARMA líkönin svo vel).
12. Vegna þess hve long-memory líkönin eru erfið hafa ýmsir höfundar stungið upp á ýmsum nálgunum. Sumt hefur verið forritað í R.
13. Eru ensku gögnin góð? Ég valdi þetta því raðirnar eru langar. Ég held að þær sé svolítið möndlaðar, urban-warming hefur verið dregin frá í mælingum eftir 1974 og meðaltöl reiknuð yfir svæði. Prívat treysti ég gögnum frá Eyrarbakka og Stykkishólmi betur.