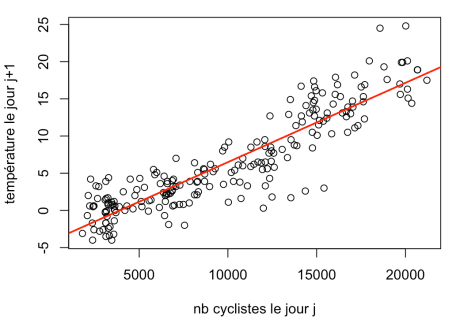
A few days ago, I came back on a sentence I found (in a French newspaper), where someone was claiming that

“… an old variable explains 85% of the change in a new variable. So we can talk about causality”

and when it was explainedthat it was just stupid : if we consider the regression of the temperature on day t+1 against the number of cyclist on day t, the R^2 exceeds 80%… but it is hard to claim that the number of cyclists on specific day will actually **cause** the temperature on the next day…



Nevertheless, that was frustrating, and I was wondering if there was a clever way to test for causality in that case. A popular one is [Granger causality](https://en.wikipedia.org/wiki/Granger_causality) (I can mention a paper we published a few years ago where we use such a test, [Tents, Tweets, and Events: The Interplay Between Ongoing Protests and Social Media](https://onlinelibrary.wiley.com/doi/abs/10.1111/jcom.12145)). To explain that test, consider a bivariate time series (just like the one we have here), \boldsymbol{z}\_t=(x\_t,y\_t), and consider some bivariate autoregressive model  
{\displaystyle {\begin{bmatrix}x\_{t}\\y\_{t}\end{bmatrix}}={\begin{bmatrix}c\_{1}\\c\_{2}\end{bmatrix}}+{\begin{bmatrix}a\_{1,1}&\textcolor{red}{a\_{1,2}}\\\textcolor{blue}{a\_{2,1}}&a\_{2,2}\end{bmatrix}}{\begin{bmatrix}x\_{t-1}\\y\_{t-1}\end{bmatrix}}+{\begin{bmatrix}u\_{t}\\v\_{t}\end{bmatrix}}}where \boldsymbol{\varepsilon}\_t=(u\_t,v\_t) is some bivariate white noise, in the sense that (i) {\displaystyle \mathbb{E} (\boldsymbol{\varepsilon}\_{t})=\boldsymbol{0}} (the noise is centered) (ii) {\displaystyle \mathbb{E} (\boldsymbol{\varepsilon}\_{t}\boldsymbol{\varepsilon}\_{t}^\top)=\Omega } , so the variance matrix is constant, but possibly non-diagonal (iii) {\displaystyle \mathbb{E} (\boldsymbol{\varepsilon}\_{t}\boldsymbol{\varepsilon}\_{t-h}^\top)=\boldsymbol{0} } for all h\neq 0. Note that we can use the simplified expression{\displaystyle {\boldsymbol{z}\_t=\boldsymbol{c}+\boldsymbol{A}\boldsymbol{z}\_{t-1}+\boldsymbol{\varepsilon}\_t}}Now, Granger test is based on several quantities. With off-diagonal terms of matrix \Omega, we have a so-called *instantaneous* causality, and since \Omega is symmetry, we will write x\leftrightarrow y. With off-diagonal terms of matrix \boldsymbol{A}, we have a so-called *lagged* causality, with either \textcolor{blue}{x\rightarrow y} or \textcolor{red}{x\leftarrow y} (and possibly both, if both terms are significant).

So I wanted to try on my two-variable problem.

|  |  |
| --- | --- |
| 1  2  3  4 | **df** = **read.csv**("cyclistsTempHKI.csv")  dfts = **cbind**(**C**=**ts**(**df**$cyclists,**start** = **c**(2014, 1,2), **frequency** = 365),  **T**=**ts**(**df**$meanTemp,**start** = **c**(2014, 1,2), **frequency** = 365))  **library**(vars) |

I now have “time series” objects, and we can fit a VAR model,

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13 | var2 = VAR(dfts, p = 1, type = "const")  **coefficients**(var2)  $C  Estimate Std. Error **t** value Pr(&gt;|t|)  C.l1 0.8684009 0.02889424 30.054460 8.080226e-107  T.l1 70.3042012 20.07247411 3.502518 5.102094e-04  const 807.6394001 187.75472482 4.301566 2.110412e-05    $T  Estimate Std. Error **t** value Pr(&gt;|t|)  C.l1 0.0003865391 6.257596e-05 6.177118 1.540467e-09  T.l1 0.6611135594 4.347074e-02 15.208241 6.086394e-42  const -1.6413074565 4.066184e-01 -4.036481 6.446018e-05 |

For instant, we can run a causality, to test if the number of cyclists can cause the temperature (on the next day)

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | causality(var2, cause = "C")  $Granger    Granger causality H0: **C** do not Granger-cause **T**    **data**: VAR object var2  F-Test = 38.157, df1 = 1, df2 = 842, p-value = 1.015e-09 |

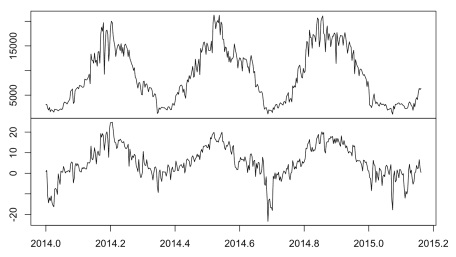
Here, we should clearly reject H\_0, which is that there is no causal effect. Which is the way statistician say that there should be some causal effect between the number of cyclist and the temperature…

So clearly, something is wrong here. Either it is some sort of superpower that cyclists are not aware of. Or this test that was used for [forty years](https://linkinghub.elsevier.com/retrieve/pii/016518898090069X) (Clive Granger even got a Nobel price for it) is not working. Or we missed something. Actually… I think we missed something here. The series are not stationary. We can almost see it with

|  |  |
| --- | --- |
| 1  2  3  4  5 | Phi = **matrix**(**c**(**coefficients**(var2)$C[1:2,1],**coefficients**(var2)$T[1:2,1]),2,2)  **eigen**(Phi)  **eigen**() decomposition  $values  [1] 0.9594810 0.5700335 |

where the highest eigenvalue is very close to one. But actually, we look here at the temperature…

|  |  |
| --- | --- |
| 1 | **plot**(dfts) |



i.e. at least, we should expect some seasonal unit root here. So let us use two techniques. The first one is a classical one-year difference, \Delta\_{365}\boldsymbol{z}\_t=\boldsymbol{z}\_t-\boldsymbol{z}\_{t-365}

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13 | var2 = VAR(**diff**(dfts,365), p = 1, type = "const")  **coefficients**(var2)  $C  Estimate Std. Error **t** value Pr(&gt;|t|)  C.l1 0.8376424 0.07259969 11.537823 1.993355e-16  T.l1 42.2638410 28.58783276 1.478386 1.449076e-01  const -507.5514795 219.40240747 -2.313336 2.440042e-02    $T  Estimate Std. Error **t** value Pr(&gt;|t|)  C.l1 0.000518209 0.0003277295 1.5812096 1.194623e-01  T.l1 0.598425288 0.1290511945 4.6371154 2.162476e-05  const 0.547828079 0.9904263469 0.5531235 5.823804e-01 |

The test on the fited VAR model yields

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | causality(var2, cause = "C")  $Granger    Granger causality H0: **C** do not Granger-cause **T**    **data**: VAR object var2  F-Test = 2.5002, df1 = 1, df2 = 112, p-value = 0.1167 |

i.e., with a 11% p-value, we should reject the assumption that the number of cyclists cause the temperature (on the next day), and actually, we should also reject the other way

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | causality(var2, cause = "T")  $Granger    Granger causality H0: **T** do not Granger-cause **C**    **data**: VAR object var2  F-Test = 2.1856, df1 = 1, df2 = 112, p-value = 0.1421 |

Nevertheless, if we look at the instantaneous causality, this one makes more sense

|  |  |
| --- | --- |
| 1  2  3  4  5  6 | $Instant    H0: No instantaneous causality between: **T** and **C**    **data**: VAR object var2  Chi-squared = 13.081, **df** = 1, p-value = 0.0002982 |

The second idea would be to use a one day difference, \Delta\_{1}\boldsymbol{z}\_t=\boldsymbol{z}\_t-\boldsymbol{z}\_{t-1} and to fit a VAR model on that one

|  |  |
| --- | --- |
| 1  2  3  4 | VARselect(**diff**(dfts,1), lag.max = 4, type="const")  $selection  **AIC**(n) HQ(n) SC(n) FPE(n)  3 3 2 3 |

but on that one, a VAR(1) model – with only one lag – might not be sufficient. It might be better to consider a VAR(3)

|  |  |
| --- | --- |
| 1 | var2 = VAR(**diff**(dfts,1), p = 3, type = "const") |

and on that one, one more time, we should reject the causal effect of the number of cyclists on the temperature (on the next day)

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7 | causality(var2, cause = "C")  $Granger    Granger causality H0: **C** do not Granger-cause **T**    **data**: VAR object var2  F-Test = 0.67644, df1 = 3, df2 = 828, p-value = 0.5666 |

and this time, there could be a (lagged) causal effect of the temperature on the number of cyclists

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14 | causality(var2, cause = "T")  $Granger    Granger causality H0: **T** do not Granger-cause **C**    **data**: VAR object var2  F-Test = 7.7981, df1 = 3, df2 = 828, p-value = 3.879e-05    $Instant    H0: No instantaneous causality between: **T** and **C**    **data**: VAR object var2  Chi-squared = 55.83, **df** = 1, p-value = 7.905e-14 |

but nothing instantaneously… So it looks like Granger causality performs well on that one !