Man glätte niemals Zeitreihen!

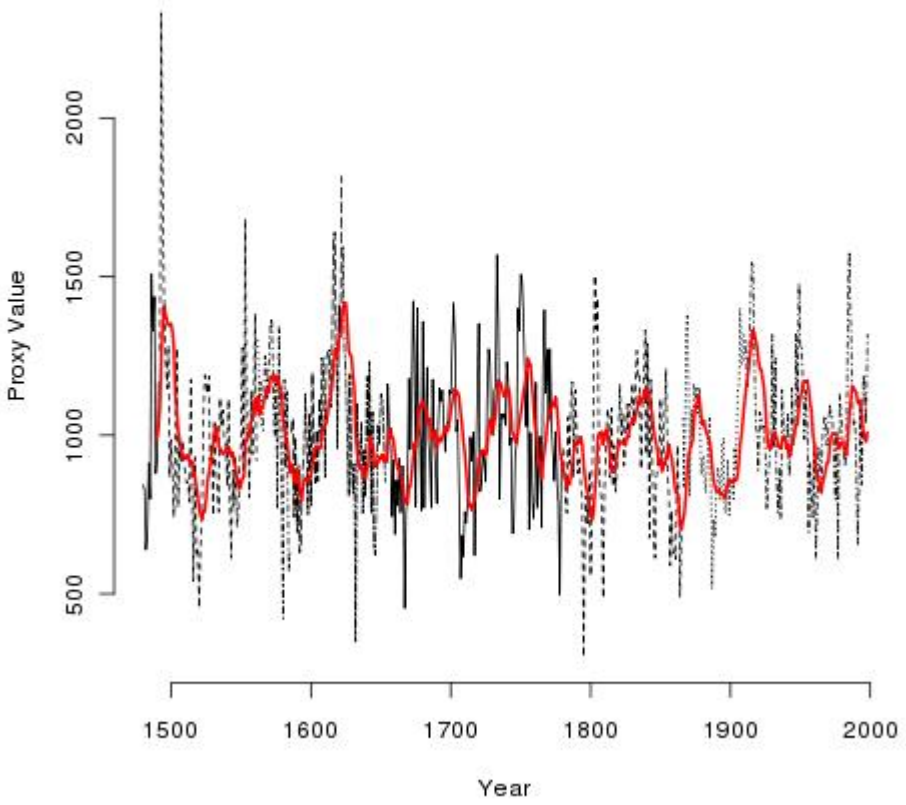
geschrieben von Chris Frey | 8. Januar 2021

Der Originaltitel ist als Wortspiel nur schwer übersetzbar: ***Do not smooth times series, you hockey puck!***

Der Ratschlag, der den Titel dieses Beitrags bildet, wäre der eines Statistikers, wie man keine Zeitreihenanalyse durchführt. Nach den Methoden zu urteilen, die ich regelmäßig auf Daten dieser Art angewendet sehe, ist diese Zurechtweisung dringend nötig.

Der Ratschlag ist jetzt besonders relevant, weil sich eine neue Hockeystick-Kontroverse zusammenbraut. Mann und andere haben eine neue Studie veröffentlicht, in der viele Daten zusammengeführt wurden, und sie behaupten, erneut gezeigt zu haben, dass das Hier und Jetzt heißer ist als das Damals und Dort. Man gehe zu climateaudit.org und lese alles darüber. Ich kann es nicht besser machen als Steve, also werde ich es nicht versuchen. Was ich tun kann, ist zu zeigen, wie man es nicht tun soll. Ich werde es auch schreien, denn ich möchte sicher sein, dass jeder es hört.

Mann stellt auf dieser Site eine große Anzahl von Temperatur-Proxy-Datenreihen zur Verfügung. Hier ist eine von ihnen mit der Bezeichnung *wy026.ppd* (ich habe einfach eine aus dem Haufen herausgegriffen). Hier ist das Bild dieser Daten:



Die verschiedenen schwarzen Linien sind die tatsächlichen Daten! Die rote Linie ist ein geglätteter 10-Jahres-Mittelwert! Ich nenne die schwarzen Daten die realen Daten, und die geglätteten Daten die fiktiven Daten. Mann hat einen „Tiefpassfilter“ verwendet, der sich vom laufenden Mittelwert unterscheidet, um seine fiktiven Daten zu erzeugen, aber eine Glättung ist eine Glättung, und was ich jetzt sage, ändert sich kein bisschen, je nachdem, welche Glättung man verwendet.

> Jetzt werde ich die große Wahrheit der Zeitreihenanalyse verkünden. Solange die Daten nicht mit Fehlern gemessen werden, ***glätte*** **man nie, niemals, aus keinem Grund, unter keiner Drohung, die Reihe!** Und wenn man sie aus irgendeinem bizarren Grund doch glättet, **verwende man die geglättete Reihe AUF KEINEN FALL als Input für andere Analysen!** Wenn die Daten mit Fehlern gemessen werden, kann man versuchen, sie zu modellieren (was bedeutet, sie zu glätten), um den Messfehler abzuschätzen, aber selbst in diesen seltenen Fällen muss man eine externe (das gelehrte Wort ist „exogene“) Schätzung dieses Fehlers haben, d.h. eine, die nicht auf den aktuellen Daten basiert.

[Alle Hervorhebungen im Original]

Wenn man in einem Moment des Wahnsinns Zeitreihendaten glättet und sie als Eingabe für andere Analysen verwendet, erhöht man dramatisch die Wahrscheinlichkeit, sich selbst zu täuschen! Das liegt daran, dass die Glättung Störsignale hervorruft – Signale, die für andere Analysemethoden echt aussehen. Egal wie, man wird sich seiner Endergebnisse zu sicher sein! Mann et al. haben ihre Reihen erst dramatisch geglättet und dann separat analysiert. Unabhängig davon, ob ihre These stimmt – ob es wirklich einen dramatischen Temperaturanstieg in letzter Zeit gibt – sind sie sich ihrer Schlussfolgerung nun garantiert zu sicher.

Und jetzt zu einigen Details:

● Ein Wahrscheinlichkeitsmodell sollte nur für eine Sache verwendet werden: um die Unsicherheit von **noch nicht gesehenen Daten** zu quantifizieren. Ich gehe immer wieder darauf ein, weil diese einfache Tatsache aus unerfindlichen Gründen offenbar schwer zu merken ist.

● Die logische Folge dieser Wahrheit ist, dass die Daten in einer Zeitreihenanalyse die Daten sind. Diese Tautologie ist dazu da, um zum Nachdenken anzuregen. Die Daten sind die Daten! Die Daten sind nicht irgendein Modell derselben. Die realen, tatsächlichen Daten sind die realen, tatsächlichen Daten. Es gibt keinen geheimen, versteckten „zugrundeliegenden Prozess“, den man mit irgendeiner statistischen Methode herauskitzeln kann und der die „echten Daten“ zeigen wird. Wir kennen die Daten bereits und sie sind da. Wir glätten sie nicht, um uns zu sagen, was es „wirklich ist“, weil wir bereits wissen, was es „wirklich ist“.

● Es gibt also nur zwei Gründe (abgesehen von Messfehlern), jemals Zeitreihendaten zu modellieren:

1. Um die Zeitreihe mit externen Faktoren in Verbindung zu bringen. Dies ist das Standard-Paradigma für 99 % aller statistischen Analysen. Man nehme mehrere Variablen und versuche, die Korrelation usw. zu quantifizieren, aber nur mit dem Gedanken, den nächsten Schritt zu tun.

2. Um zukünftige Daten vorherzusagen. **Wir brauchen die Daten, die wir bereits haben, nicht vorherzusagen**. Wir können nur vorhersagen, was wir nicht wissen, nämlich zukünftige Daten. So brauchen wir die Baumring-Proxydaten nicht vorherzusagen, weil wir sie bereits kennen.

● **Die Baumringdaten sind nicht die Temperatur!** Deshalb werden sie Proxy-Daten genannt. Ist es ein perfekter Proxy? War die letzte Frage eine rhetorische Frage? War das auch eine? Weil es ein Proxy ist, muss die Unsicherheit seiner Fähigkeit, die Temperatur vorherzusagen, in den Endergebnissen berücksichtigt werden. Hat Mann das getan? Und was genau ist eine rhetorische Frage?

● Es gibt Hunderte von Zeitreihen-Analysemethoden, die meisten mit dem Ziel, die Unsicherheit des Prozesses zu verstehen, damit zukünftige Daten vorhergesagt werden können und die Unsicherheit dieser Vorhersagen quantifiziert werden kann (dies ist aus gutem Grund ein riesiges Studiengebiet, z. B. auf den Finanzmärkten). Dies ist eine legitime Verwendung von Glättung und Modellierung.

● Wir sollten sicherlich die Beziehung zwischen dem Proxy und der Temperatur modellieren und dabei die sich im Laufe der Zeit verändernde Natur des Proxys berücksichtigen, die unterschiedlichen physikalischen Prozesse, die dazu führen, dass sich der Proxy unabhängig von der Temperatur verändert, oder wie die Temperatur diese Prozesse verstärkt oder auslöscht, und so weiter und so fort. Aber wir sollten nicht damit aufhören, wie es alle getan haben, etwas über die Parameter der Wahrscheinlichkeitsmodelle zu sagen, die zur Quantifizierung dieser Beziehungen verwendet werden. Dadurch wird man sich der Endergebnisse wieder einmal *viel zu sicher*. Uns interessiert nicht, wie der Proxy die mittlere Temperatur vorhersagt, uns *interessiert*, wie der Proxy die *Temperatur* vorhersagt.

● Wir brauchen keinen statistischen Test, um zu sagen, ob eine bestimmte Zeitreihe seit einem bestimmten Zeitpunkt gestiegen ist. Warum? Wenn man es nicht weiß, gehe man zurück und lese diese Punkte von Anfang an. Es liegt daran, dass wir uns nur die Daten ansehen müssen: wenn sie einen Anstieg zeigen, dürfen wir sagen: „Sie [die Zeitreihe] hat zugenommen.“ Wenn sie nicht gestiegen sind oder gar abgenommen haben, dann dürfen wir *nicht* sagen: „sie hat zugenommen.“ So einfach ist es wirklich.

● Man kann mir jetzt sagen: „OK, Herr Neunmalklug. Was wäre, wenn wir mehrere verschiedene Zeitreihen von verschiedenen Orten hätten? Wie können wir feststellen, ob es einen generellen Anstieg bei allen gibt? Wir brauchen sicherlich Statistiken und p-Werte und Monte-Carlo-Berechnungen, um uns zu sagen, dass sie zugenommen haben oder dass die ‚Nullhypothese‘ von keiner Zunahme wahr ist.“ Erstens hat mich niemand schon lange „Herr Neunmalklug“ genannt, also sollten Sie sich Ihre Sprache besser überlegen. Zweitens: Haben Sie nicht aufgepasst? Wenn Sie sagen wollen, dass 52 von 413

Zeitreihen seit einem bestimmten Zeitpunkt gestiegen sind, dann schauen Sie sich die Zeitreihen an und zählen Sie! Wenn 52 von 413 Zeitreihen gestiegen sind, dann können Sie sagen „52 von 413 Zeitreihen sind gestiegen.“ Wenn mehr oder weniger als 52 von 413 Zeitreihen gestiegen sind, dann können Sie nicht sagen, dass „52 von 413 Zeitreihen gestiegen sind.“ Sie können es zwar sagen, aber Sie würden lügen. Es gibt absolut keinen Grund, über Nullhypothesen usw. zu schwätzen.

Wenn Ihnen die Punkte – es ist wirklich nur ein Punkt – die ich anspreche, langweilig erscheinen, dann habe ich es geschafft. Die einzige faire Art, über vergangene, bekannte Daten in der Statistik zu sprechen, ist, sie einfach zu betrachten. Es ist wahr, dass das Betrachten von massiven Datensätzen schwierig ist und immer noch eine Art Kunst darstellt. Aber Schauen ist Schauen und es ist völlig gleichberechtigt. Wenn Sie sagen wollen, wie Ihre Daten mit anderen Daten in Beziehung standen, dann müssen Sie wiederum nur schauen.

Der einzige Grund, ein statistisches Modell zu erstellen, ist die Vorhersage von Daten, die man nicht gesehen hat. Im Fall der Proxy-/Temperaturdaten haben wir die Proxies, aber wir haben nicht die Temperatur, so dass wir sicherlich ein Wahrscheinlichkeitsmodell verwenden können, um unsere Unsicherheit in Bezug auf die nicht gesehenen Temperaturen zu quantifizieren. Aber wir können diese Modelle nur erstellen, wenn wir gleichzeitige Messungen der Proxies und der Temperatur haben. Nachdem diese Modelle erstellt sind, gehen wir wieder zu dem Punkt zurück, an dem wir die Temperatur nicht haben, und können sie vorhersagen (wobei wir daran denken müssen, dass wir nicht ihren Mittelwert, sondern die *tatsächlichen Werte* vorhersagen müssen; außerdem müssen wir berücksichtigen, wie die Beziehung zwischen Temperatur und Proxy in der Vergangenheit anders gewesen sein könnte, und wie die anderen vorhandenen Bedingungen diese Beziehung verändert haben könnten, und so weiter und so fort).

Was man nicht tun kann oder sollte ist, zuerst die Proxydaten zu modellieren/glätten, um fiktive Daten zu erzeugen und dann zu versuchen, die fiktiven Daten und die Temperatur zu modellieren. Dieser Trick wird einen immer – einfach immer – zu sicher machen und in die Irre führen. Man beachte, wie die gelesenen fiktiven Daten viel strukturierter aussehen als die realen Daten und es wird verständlich.

Der nächste Schritt ist, mit den Proxydaten selbst zu spielen und zu sehen, was zu sehen ist. Sobald mir der Wunsch erfüllt wird,

jeden Tag mit 48 Stunden zu füllen, werde ich das tun können.

Link: https://wmbriggs.com/post/195/

Übersetzt von Chris Frey EIKE