

日本の年平均気温図を作ってみた話

5th December 2020, Japan.R 2020
Yuta Kanzawa @yutakanzawa

Data Science Senior Analyst at Janssen Pharmaceutical K.K., Tokyo
A Family Company of Johnson & Johnson



I am...

- 神沢雄大 **Yuta Kanzawa** (twitter: [@yutakanzawa](https://twitter.com/@yutakanzawa))
- Data scientist at **Janssen Japan**, Tokyo
 - A pharmaceutical company of **J&J**
- Opera & wine lover
 - Wagner
 - Bourgogne
- 7 languages
 - Human: Japanese, English, German
 - Computer: R, Python, SAS, SQL



アジェンダ

- 今日話すこと
 - データ
 - モデリング
 - 推定結果
- 今日話さないこと
 - Python

TL;DR

**データ分析が好きなら
好きなものを分析してみるといいよ！**

前回までの話：Tokyo.R

- ・国土数値情報APIのススメ
- ・日本の気象観測データの（オープンなようでオープンでない）話

Previously on Tokyo.R #84 & 86

始まり：日本の1キロメッシュの年平均気温図を作りたい。



The screenshot shows the R-bloggers website interface. At the top is the R-bloggers logo and tagline. Below is a navigation bar with links like Home, About, RSS, add your blog!, Learn R, R jobs, and Contact us. The main content area features an article titled "Temperature changes in Germany visualized in R" by Yongfu, Liao, dated November 5, 2019. The article includes social media share buttons (Like, Share, Tweet, LinkedIn) and a text box for sharing content. A sidebar on the left contains a "WELCOME!" message, a "Follow @rbloggers" button, and a "Subscribe" button with a reader count of 52,206. A "Most Visited Articles of the Week" section lists nine articles, including "Tidying the John Hopkins Covid-19 data" and "How to create a simple Coronavirus dashboard specific to your country in R".



The screenshot shows the ZEIT ONLINE website. The header includes the ZEIT ONLINE logo and a search bar. Below the header is a navigation bar with categories like Politik, Gesellschaft, Wirtschaft, Kultur, Wissen, Digital, Campus, Arbeit, Entdecken, Sport, ZEITmagazin, Podcasts, and mehr. A "SCHWERPUNKTE" section highlights "DONALD TRUMP", "KLIMAGIPFEL", and "UKRAINE-KONFLIKT". The main article is titled "Klimawandel: Viel zu warm hier" and features a large grid of colorful, abstract shapes representing temperature data. The text below the grid states: "Durch den Klimawandel hat sich Deutschland deutlich erwärmt. Wir zeigen, wie sich die Temperatur in Ihrer Gemeinde in den letzten 137 Jahren entwickelt hat." The article is attributed to Paul Blicke, Elena Erdmann, and Flavio Dortana, with 145 comments.

*1 <https://www.r-bloggers.com/temperature-changes-in-germany-visualized-in-r/>

*2 <https://www.zeit.de/wissen/umwelt/2019-12/klimawandel-globale-erwaermung-warming-stripes-wohnort>

元ネタを調査

- ドイツは気象局謹製の公式データがある。
 - 1881年以降の年次データ
 - 1キロメッシュ
 - Ersi ASCII grid
 - R bloggersはそのデータを使用。 → 日本では？
- そもそも、ドイツの気象局はどのようにして1キロメッシュで推定したのか？
 - 元データの添付文書を読んだが、細かい手法の記載はない。 → 類似事例？

更に調査：日本のデータ

- 気象庁謹製の「メッシュ平年値図」*1
 - 1981～2010年の平年値を使って1kmメッシュで推定した平年値

- メッシュ平年値データは、以下のとおり、「国土数値情報ダウンロードサービス」(国土交通省)のページから、ダウンロードすることが可能です。
【「国土数値情報ダウンロードサービス」(国土交通省)のページ】

▶ <http://nlftp.mlit.go.jp/ksj/index.html>

【メッシュ平年値データ(修正版)のダウンロード手順】

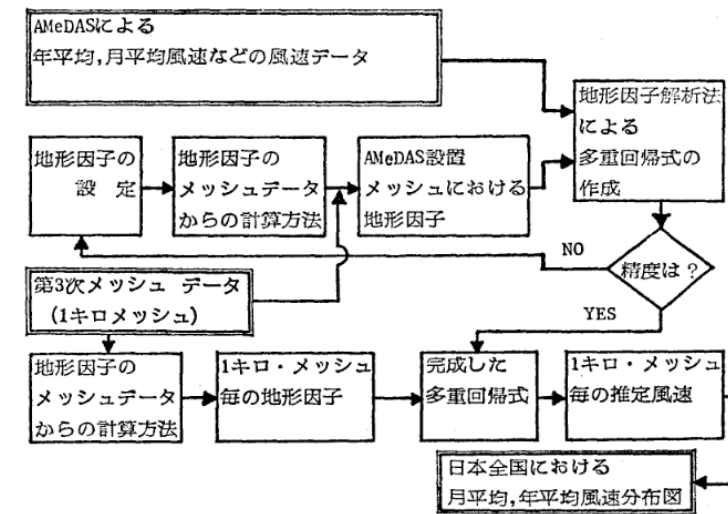
1. 上記の「国土数値情報ダウンロードサービス」(国土交通省)のページを開く。
2. データ形式は、「JPGIS2.1」をクリック(選択)する。
3. 「災害・防災」にある「平年値(気候)メッシュ」をクリック(選択)する。
4. 画面に従い、ダウンロードする地点を選択して、データを取得する。

*1 <https://www.data.jma.go.jp/obd/stats/etrn/view/atlas.html>

更に調査：類似事例

- 「メッシュデータを用いた地形因子解析法による任意の地点の平均風速の推定方法に関する研究」(小峯・村上・柴田・松野 1980) *1
- 「地形因子解析法による1km四方メッシュの年最大積雪深の推定」(桜井・城 1987) *2、など
- 過去の**気象データ**と**地理データ**から**重回帰**

↓ ↓
どこにある？



*1 https://www.metsoc.jp/tenki/pdf/1980/1980_12_0849.pdf

*2 <https://ci.nii.ac.jp/naid/110008012688>

調査：データ

- 気象データ
 - 気象庁「過去の気象データ・ダウンロード」*1
 - オープンデータとは程遠い状態。
- 地理データ
 - 「標高・傾斜度3次メッシュデータ」*2
 - 東日本大震災後のものはない。
 - 「土地利用3次メッシュデータ」*3
 - 同じ地形でも、建物や植生によって気温は違うはず。
 - 例：ビル街と公園


} 国土数値情報

*1 <https://www.data.jma.go.jp/gmd/risk/obsdl/index.php>

*2 <http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-G04-a.html>

*3 <http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-L03-a.html>

(前回の) Next steps

- 気象庁の気象データダウンロードサービス
 - To 中の人：もっとオープンにしてください！
- 国土数値情報API
 - To 中の人：復活させて下さい！！！！
- 日本の年平均気温図 
 - モデリング
 - まずは単純な重回帰

データとモデル

Data & model

データ

- 地理データ

- 標高・傾斜度3次メッシュデータ（2009年）：488,300件
- 土地利用3次メッシュデータ（2014年*1）：486,900件

- 気象データ

- 観測地点データ（2020年6月時点）：1,668件
- 観測データ（2019年）：1,300件
 - 観測地点ごとの年平均気温
 - 地理データと観測地点データが全て紐づく**928件**をモデル構築に使用した。
- 緯度経度からメッシュコードへの変換：jpmesh

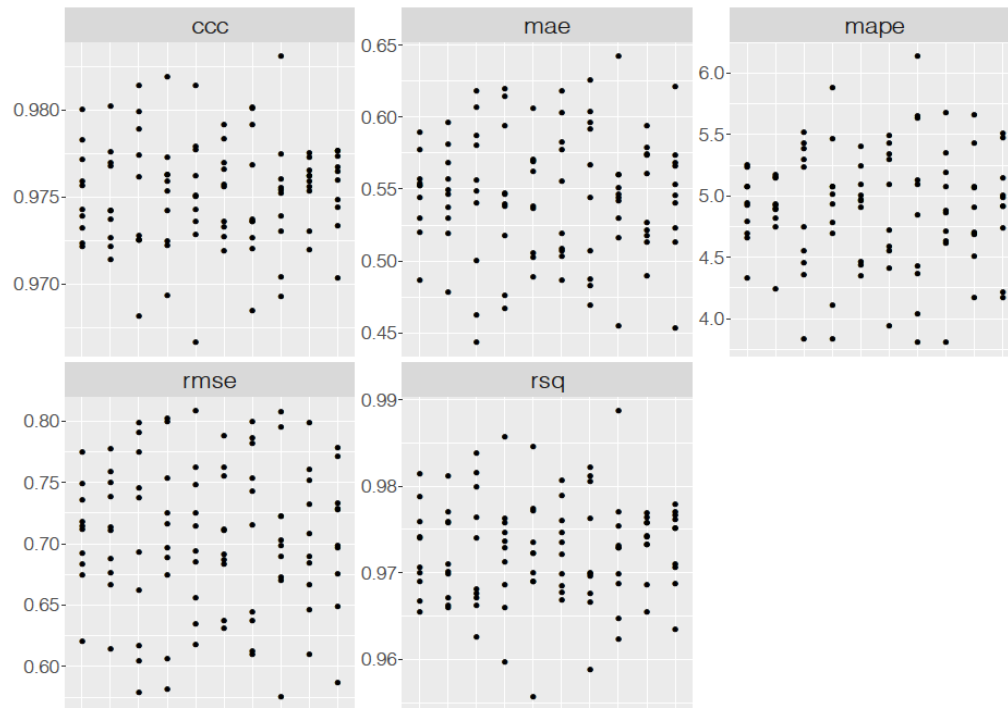
*1 2014年データになかった1次メッシュコード6740は2006年データを使用した。

変数

- 目的変数：年平均気温
- 説明変数：
 - 1kmメッシュの中心の緯度と経度
 - jpmeshで変換。
 - 地形
 - 平均標高、最高標高、最低標高、最大傾斜角度、最大傾斜方向、最小傾斜角度、最小傾斜方向、平均傾斜角度
 - 土地利用種別
 - 田、その他の農用地、森林、荒地、建物用地、道路、鉄道、その他の用地、河川地及び湖沼、海浜、海水域、ゴルフ場

線形回帰：CVで性能を推定

- 928件のモデル構築用データ
→ 10-fold CVを10回：rsample (tidymodels)



指標	平均	標準偏差
RMSE	0.705	0.0595
MAE	0.546	0.0431
MAPE	4.89	0.472
R ²	0.973	0.00589
CCC*1	0.975	0.00306

性能が良すぎて怖い...🤔 → 今後調査

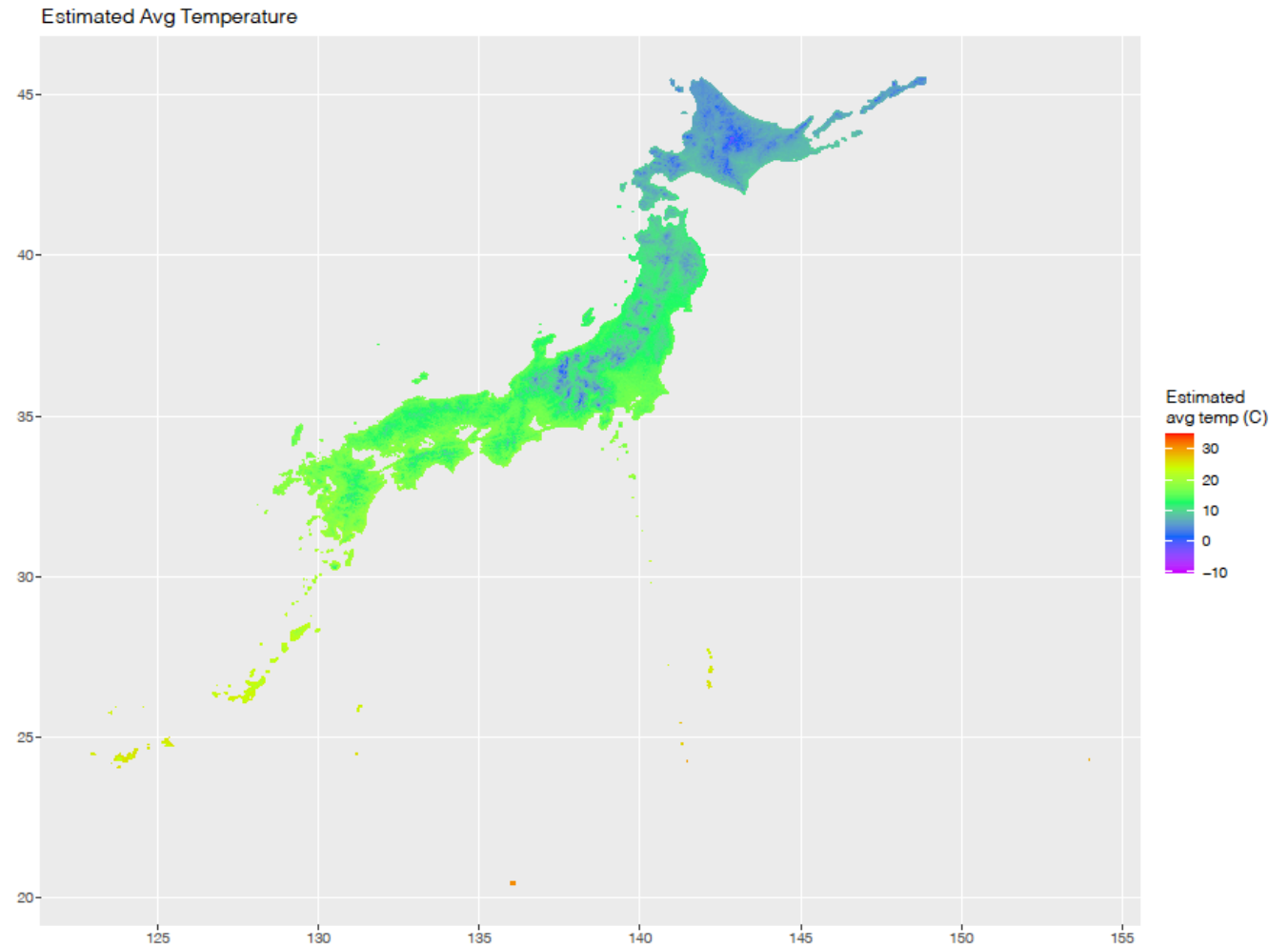
*1 Concordance Correlation Coefficient

推定結果

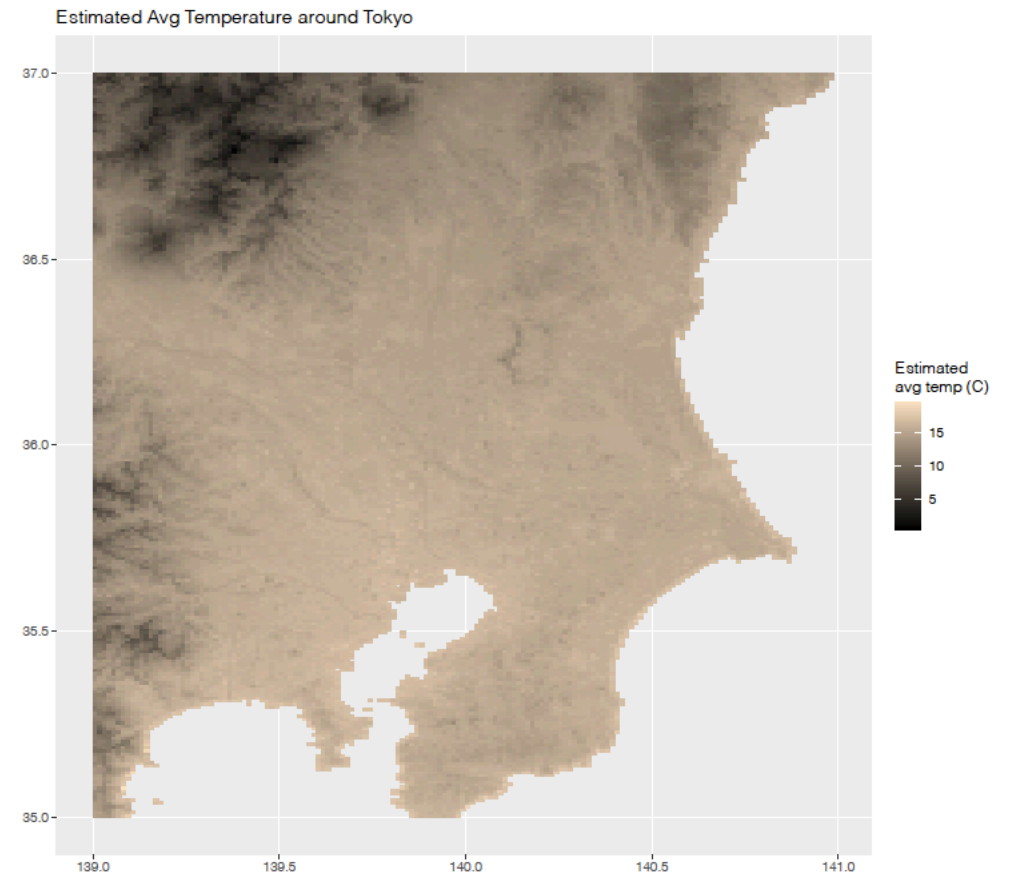
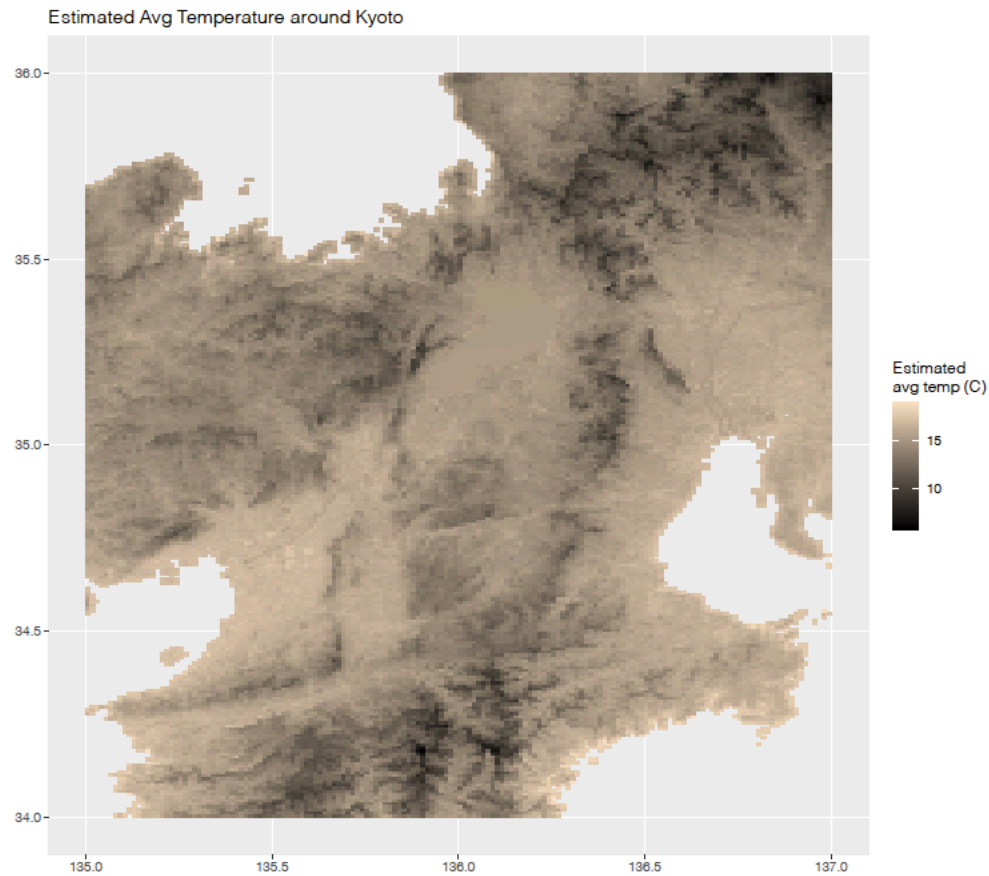
Results

日本全体（2019年）

- 経験的に違和感はない。
 - 北、山地は低い。
 - 南、低地は高い。
- 気候の影響が欠如。
 - 海流
 - 黒潮、親潮、対馬海流
 - 季節風
 - やませ
 - 日照



関西・中部と関東（グラデーションを変更）



まとめ

Long story short

Next steps

- 現行モデルの調査（精度が良すぎる...）
 - 過学習、マルチコ、リーケージ
- 推定手法の変更
 - SVM、Light GBM、深層学習
 - 状態空間モデル
- 説明変数の追加
 - 気候に関するデータ
- 時系列予測
 - 来年の年平均気温

Long story short

アヤメ*ばかりじゃつまらない。
データ分析が好きなら
好きなものを分析してみるといいよ！

免責事項：データの取得や前処理がどんなに辛くても一切責任を負いません。

* Rでは標準装備、Pythonではscikit-learnに入っている、irisデータセットのこと。

詳しくは

R Advent Calendar 2020で！

<https://qiita.com/advent-calendar/2020/rlang>



Enjoy!