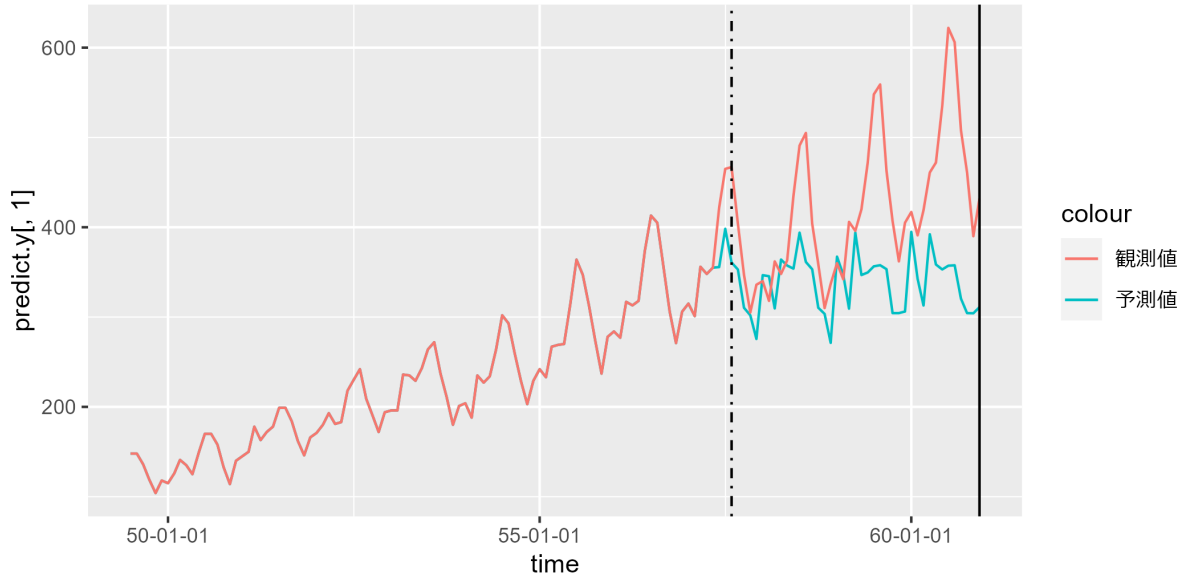
xgboostの強力な予測性能を生かした時系列予測への応用

abstract

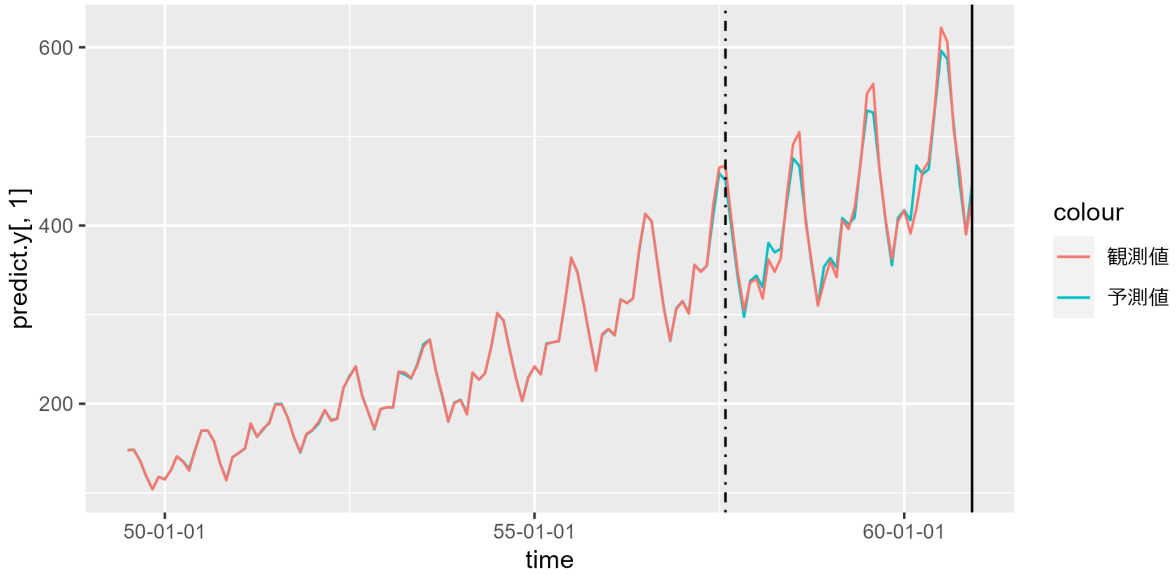
xgboost[1]は**非deep learning** の中では最も優れた予測性能を持っています。しかし、それは内挿的な問題、回帰としては旨く機能しますが外挿が必要な時系列データの未来予測に対しては旨く機能しません。データの背景にある説明変数を追加してもこの傾向は回避することは難しいと思います。可能な説明変数としてはデータから１日前、数日前等のラグ、月、日、曜日、祝日などです。実際、これらに起因するデータ値の増減は十分にありえて説明変数として機能します。しかしこれを行ったとしてもトレンドを捉える事はできません。



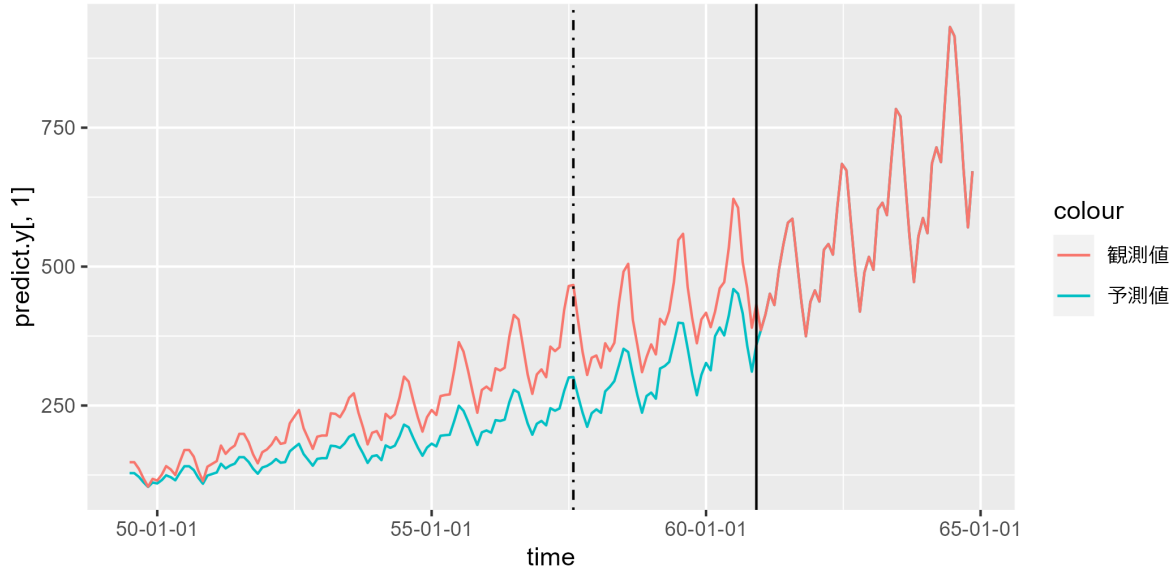
Introduction

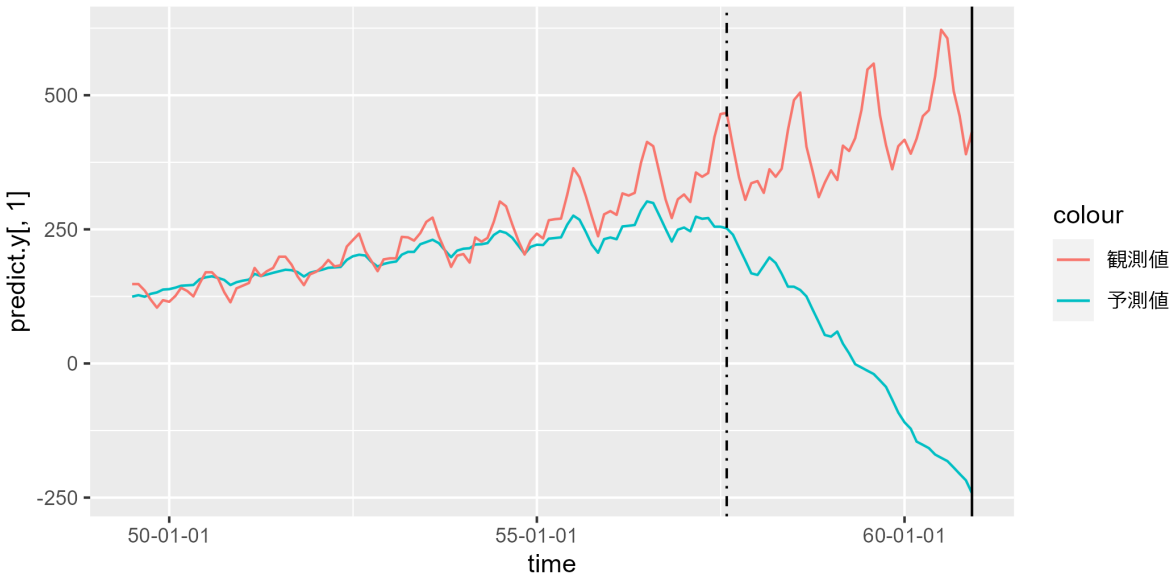
トレンドがあるような時系列データでは訓練区間ではうまく行ってもテスト区間や未来においてはうまく行きません。主な理由は簡単です。訓練した区間と全く異なる基本統計（例えば平均値）を持っているためで訓練によってまったく獲得されない情報だからです。

通常は定常性、つまりデータの平均と分散に注目するときそれはどの時点でも大体同じ傾向を示していることが必要です。従ってデータの差分や対数を取る事でデータを定常化して予測モデルを作って予測結果を差分、対数変換の逆を行うことで対処されます。



このように差分、対数変換したデータに対しては**xgboost**[1]でも一見良好な予測モデルになります。しかし、これも差分を取る事で定常になるという事がどの時点でも守られる場合に限られます。データの状況を少し変更すると実測値とのズレはとても大きくなり利用できる予測モデルにはなりません。

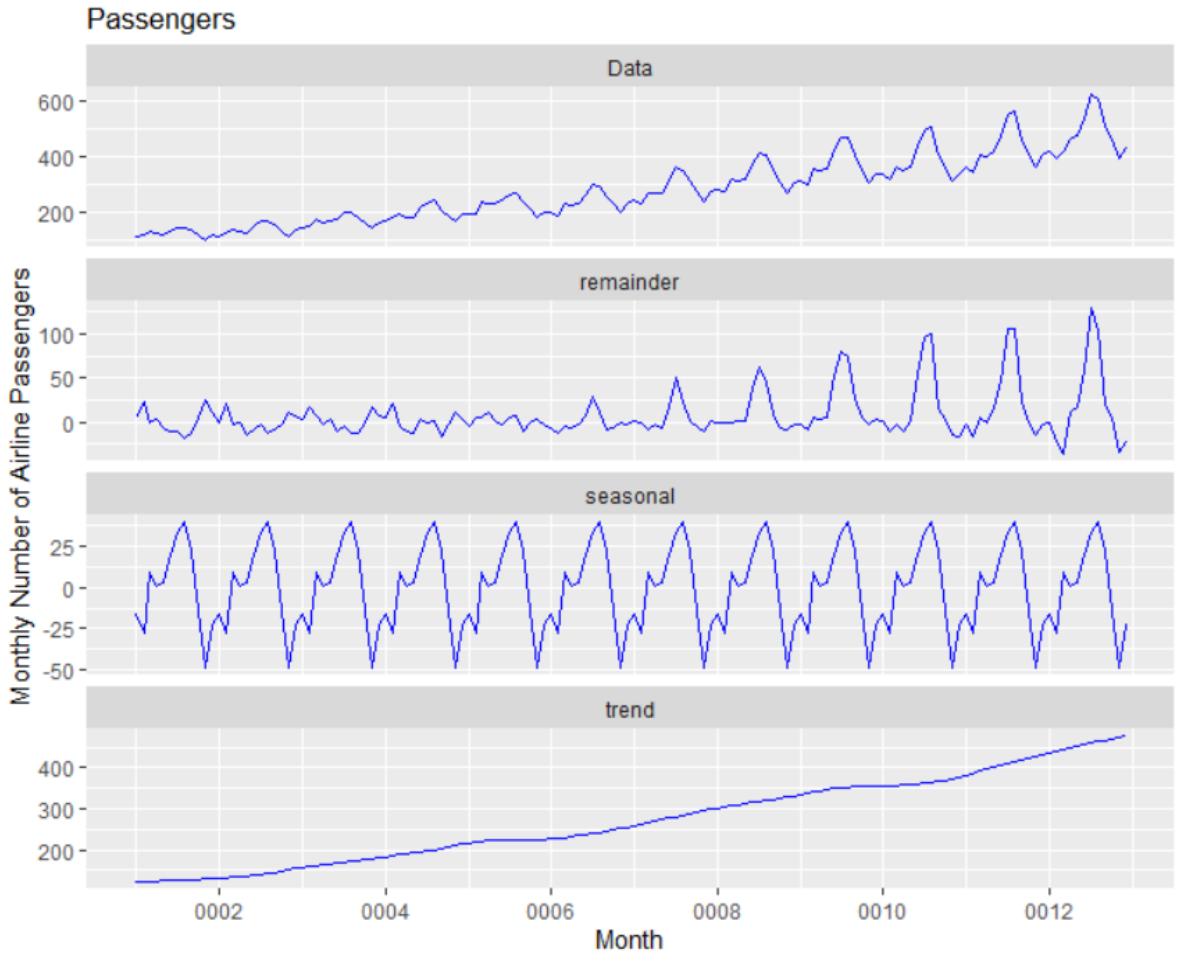




Overview of the proposed method

xgboost[1]を諦めるという事も考え方としてあります。例えばporohet[2]等は非常に優秀な結果を出力しますが説明能力に欠ける点が残ります。一方**xgboost**[1]は説明能力を持ち**非deep learning** の中では最も優れた予測性能が実証されておりそれを捨てるには非常にもったいないと思います。そこで妥協案として他のモデルと組みあわせて**xgboost**[1]の優れた部分を使うというアイデアです。

幸い時系列データはトレンドを分離することが出来ます。



データは　**trend + seasonal + remainder**　という形に分解出来ます。**xgboost**[1]には**trend**を除いた部分だけを任せて **trend** を**ARIMA**[3]等でモデル化してそれを合わせる事で優れたモデルを構築できそうです。しかし、データによっては**xgboost**[1]は繰り返される周期に対しても弱点が露見します。

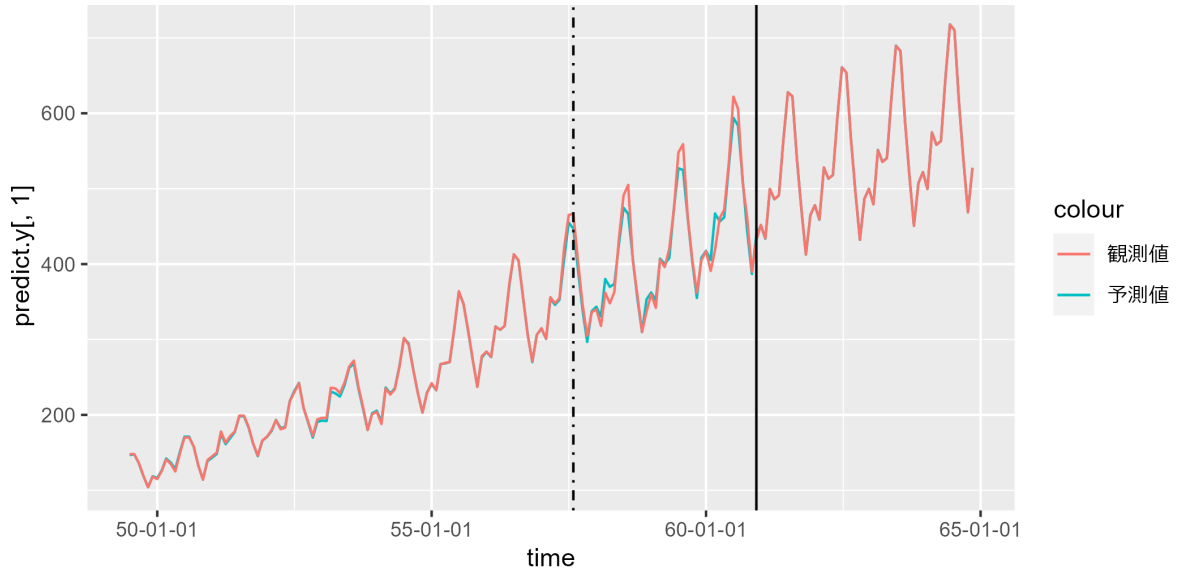
そこで、周期的成分をフーリエ展開したsin項、cos項を説明変数に追加する事で対処可能と考えています。

$$y_t = a + \sum_{k=1}^{K} \left[\alpha_k \sin(2\pi k t/m) + \beta_k \cos(2\pi k t/m) \right]$$

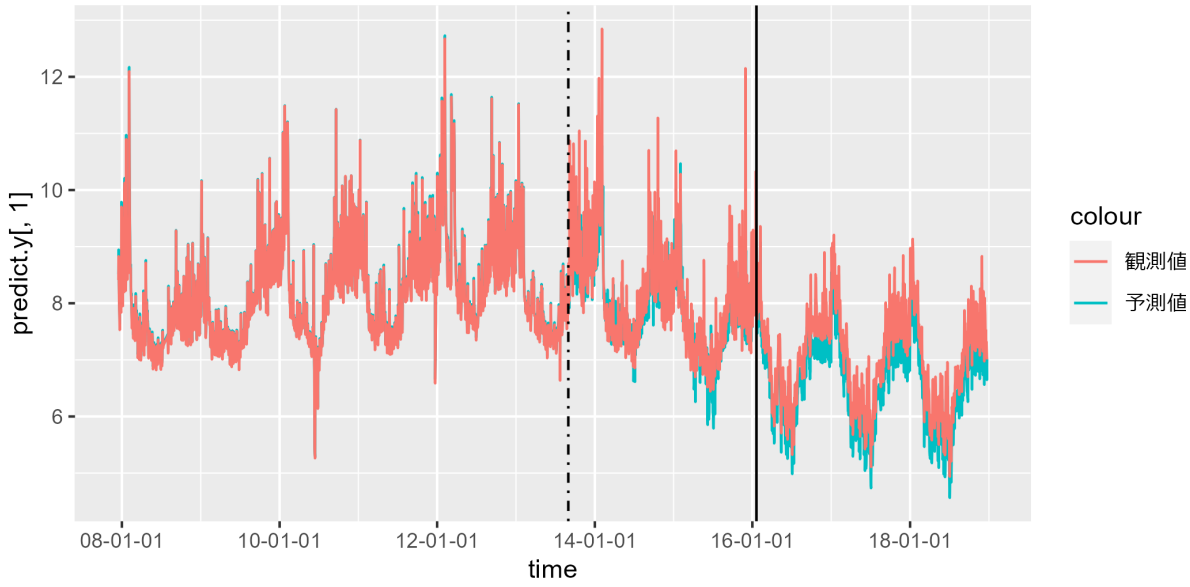
しかし、sin項、cos項が使えるの前日までに制限する必要があります。なぜなら予測したい当日のsin項、cos項は本来は未知の値になる事です。従って当日の説明変数として欠落してしまうため予測して埋めなければならなくなります。そこで私のアイデアは前日のsin項、cos項をそのまま使うという事で仮の予測を行うことが出来ます。この部分は予測が進んだ時点で再度sin項、cos項を求め直すことが出来るためこの説明変数は予測が進む毎にリフレッシュさせていく事でコピーした前日のsin項、cos項だけが続いていくという課題だけは解消出来ます。しかし、これを補強する事が可能です。時系列データを分解した結果としてseasonal が得られています。seasonal は概ね単純な繰り返しになるためこの単純な繰り返しデータを利用することでさらに妥当な説明変数に修正可能になります。

Experiment

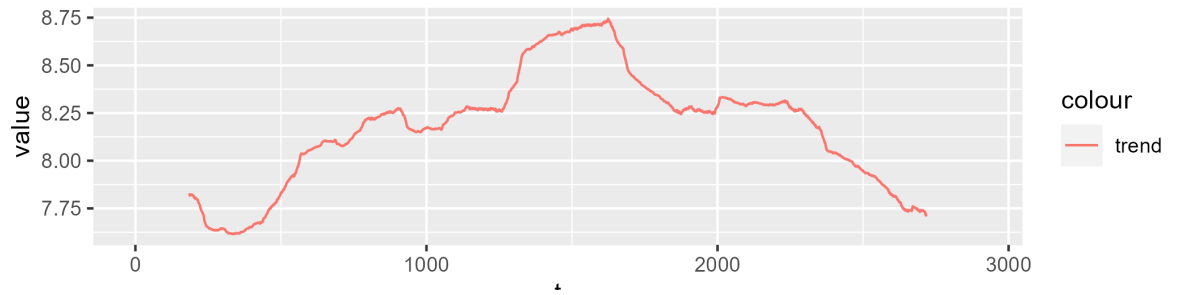
このアイデアで検証してみました。



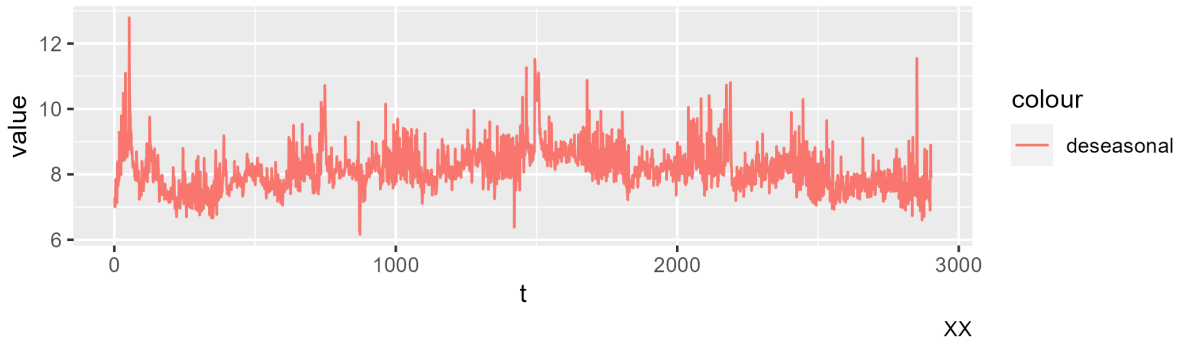
データの外つまり延長された未来に置いても妥当な予測が出来ているようです。次にexample_wp_log_peyton_manning.csvでも実験してみます。３年分（1096ステップ）を延長予測した結果です。



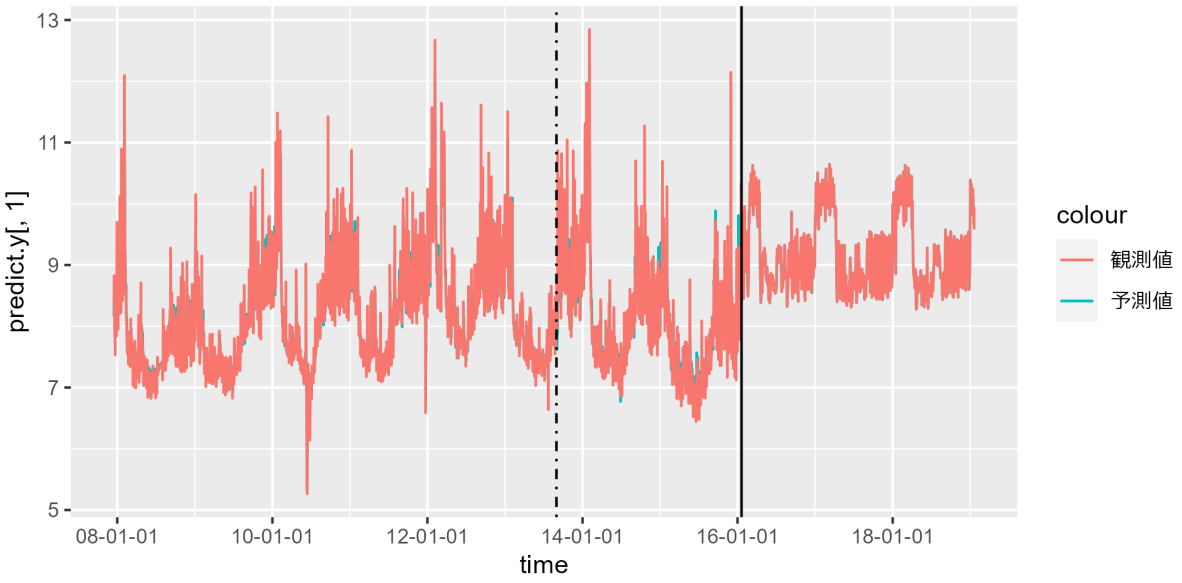
このデータのトレンド成分は



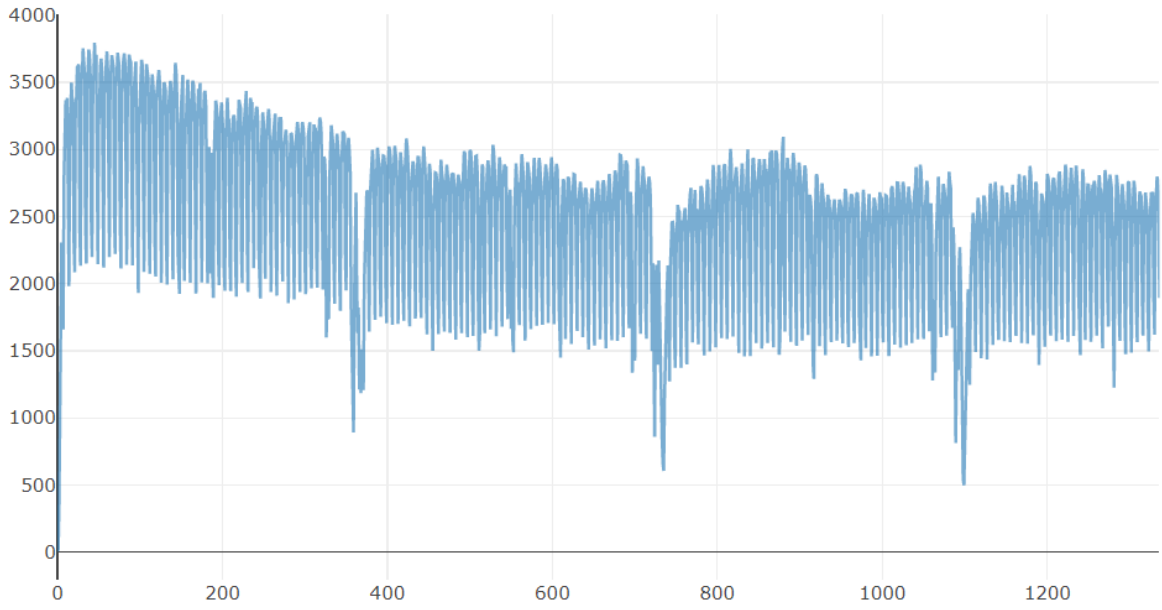
右肩下がりの傾向を示しているのでトレンドに従った予測モデルになっていることが分かります。**xgtboost**[1]が寄与した成分は下の図のようなデータになっていて**xgboost**[1]の優れた能力を十分生かした予測モデルになっています。



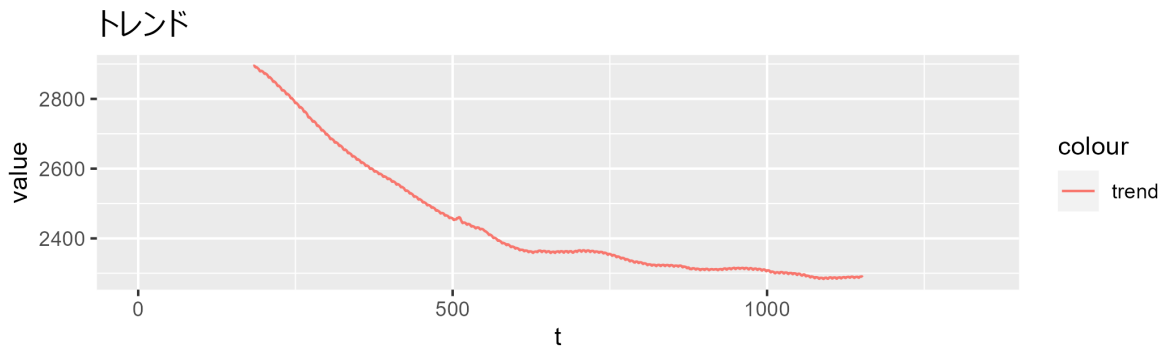
ちなみに**xgboost**[1]のみでモデル化してみると次の様になります。



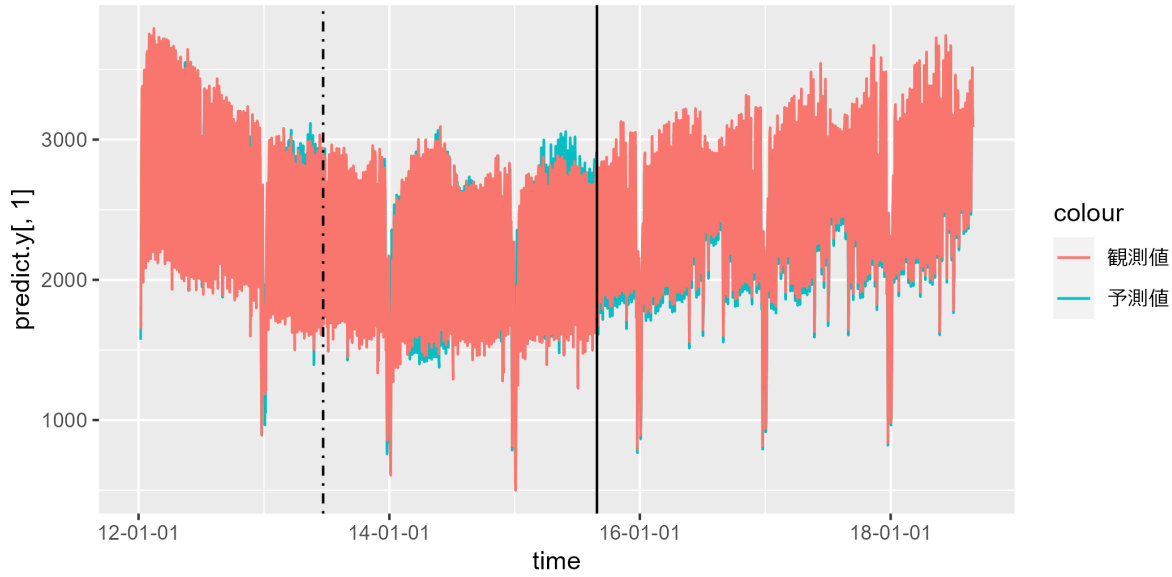
この事からもこのアイデアは妥当な結果を出力していると思われます。これ以外のデータの検証としては



このデータは右肩さがりに見えます。実際に



トレンド成分が右肩下がりになっています。このデータは訓練区間は全体の４０%に制約してトレーニングしてみました。テスト区間は非常に良い予測をしていますが３年分（1096ステップ）に延長した部分は若干右肩上がりを示してしましましたが特徴的な増減は非常に良い予測をしているように見えます。予測延長しているため実際にはわかりません。



Discussion

実験では**ARIMA**[3]を併用しましたがこの部分を**porphet**[2]を使うという事も考えられます。課題は超長期周期があるようなデータではその周期を複数現れるまでのデータが必要になる点です。その分だけ学習データ量が多くなり非deep learningの力では計算コストが非常に大きくなるという点です。

References

[1]CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016. p. 785-794.

[2]KHAYYAT, Mashael, et al. Time Series Facebook Prophet Model and Python for COVID-19 Outbreak Prediction. *CMC-COMPUTERS MATERIALS & CONTINUA*, 2021, 67.3: 3781-3793.

[3]Hyndman, RJ and Khandakar, Y (2008) "Automatic time series forecasting: The forecast package for R", *Journal of Statistical Software*, **26**(3).

Wang, X, Smith, KA, Hyndman, RJ (2006) "Characteristic-based clustering for time series data", *Data Mining and Knowledge Discovery*, **13**(3), 335-364.

PICCOLO, Domenico. A distance measure for classifying ARIMA models. *Journal of time series analysis*, 1990, 11.2: 153-164.

Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting*. Springer, New York. Sections 3.3 and 8.3.

Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

Gardner, G, Harvey, A. C. and Phillips, G. D. A. (1980). Algorithm AS 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. *Applied Statistics*, **29**, 311--322. 10.2307/2346910.

Harvey, A. C. (1993). *Time Series Models*. 2nd Edition. Harvester Wheatsheaf. Sections 3.3 and 4.4.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, **22**, 389--395. 10.2307/1268324.

Ripley, B. D. (2002) Time series in R 1.5.0. *R News*, **2/2**, 2--7. https://www.r-project.org/doc/Rnews/Rnews_2002-2.pdf

```
@inproceedings{y2021timeseriesxgboost,
  title={time series xgboost: time series xgboost},
  author={sanaxen},
  year={2021}
}
```