地理空間データの交差検証、

正しくできていますか？ (short ver.)

Shinya Uryu (u_ribo)

November 13, 2018

Tokyo.R#74LT @CyberAgent

データの自己相関

「近縁」なデータの類似性

ここでいう「近縁」

- 空間
- 時間
- 系統

… 性質が近いデータは値が類似しやすい

通常の交差検証

k-fold cross validation

- データをk個に分割
    - 分割はランダム
- kのうち一つをテストデータ、k-1個の群を訓練データとして学習
- モデルの精度検証に用いられる

そのデータで交差検証しても大丈夫?

空間データにk分割交差検証を適用する際の問題

ざっくりいうと

1. モデルエラー
    - 独立同分布の仮定に違反してしまう
2. 過学習しやすい
    - データの空間自己相関を考慮せずにランダムサンプリング

打開策

1. Spatial Cross-Validation
2. Target-oriented cross-validation

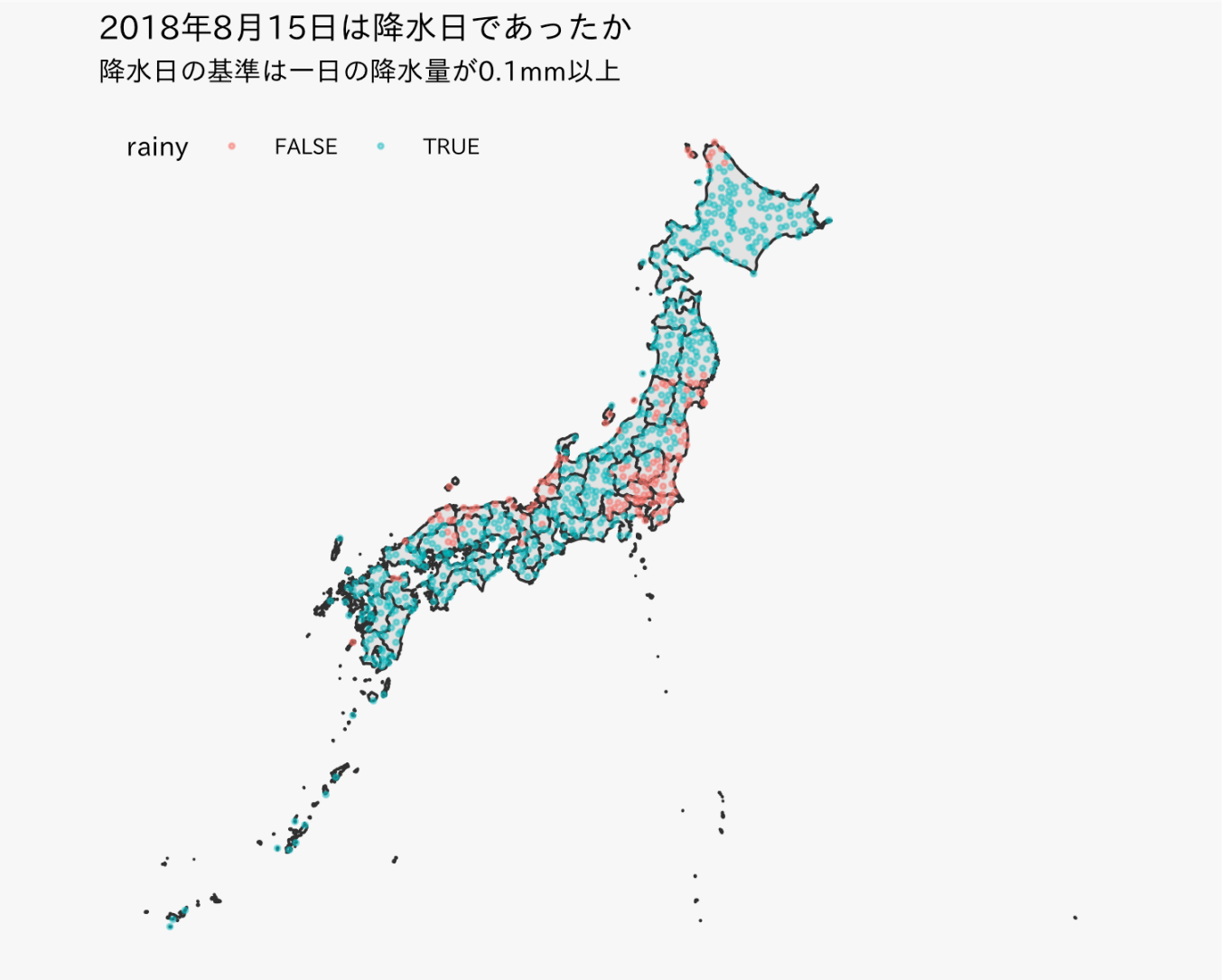
```
library(sf)
library(dplyr)
```

気象庁のデータを用いたデモ

降水量を予測するモデルを考える

- 有線ロボット気象計
  - 降水量、気温、風向、風速、日照時間
  - 今回は降水量、気温だけを利用
- 全国668地点
- 2018年8月15日
- 降水日... 一日の降水量が0.1mm以上あったかの判定

降水量を予測するモデルを考える



1/2 Spatial Cross-Validation

Step

1. 空間データの座標と他の変数を切り分ける
2. `makeClassifTask(coordinate = )`に座標データを与えたタスクを作る
3. `makeResampleDesc(method = "SpRepCV")`でデータ生成

データの用意

```
df_train <-
  df_weather_20180815 %>%
  select(elevation, temperature_mean, rainy) %>%
  st_set_geometry(NULL) %>%
  as.data.frame()

coords <-
  df_weather_20180815 %>%
  st_coordinates() %>%
  as.data.frame()
```

mlr package

```
library(mlr)

spatial.task <-
  makeClassifTask(target = "rainy",
                  data = df_train,
                  coordinates = coords,
                  positive = "TRUE")

learner.rf <-
  makeLearner("classif.ranger",
              predict.type = "prob")
```

Conventionally CV

```
resampling_cv <-
  makeResampleDesc(method = "RepCV",
                   fold = 5, reps = 5)

set.seed(123)
cv_out <-
  resample(learner    = learner.rf,
           task       = spatial.task,
           resampling = resampling_cv,
           measures   = list(auc))

mean(cv_out$measures.test$auc, na.rm = TRUE)
# [1] 0.8568344
```

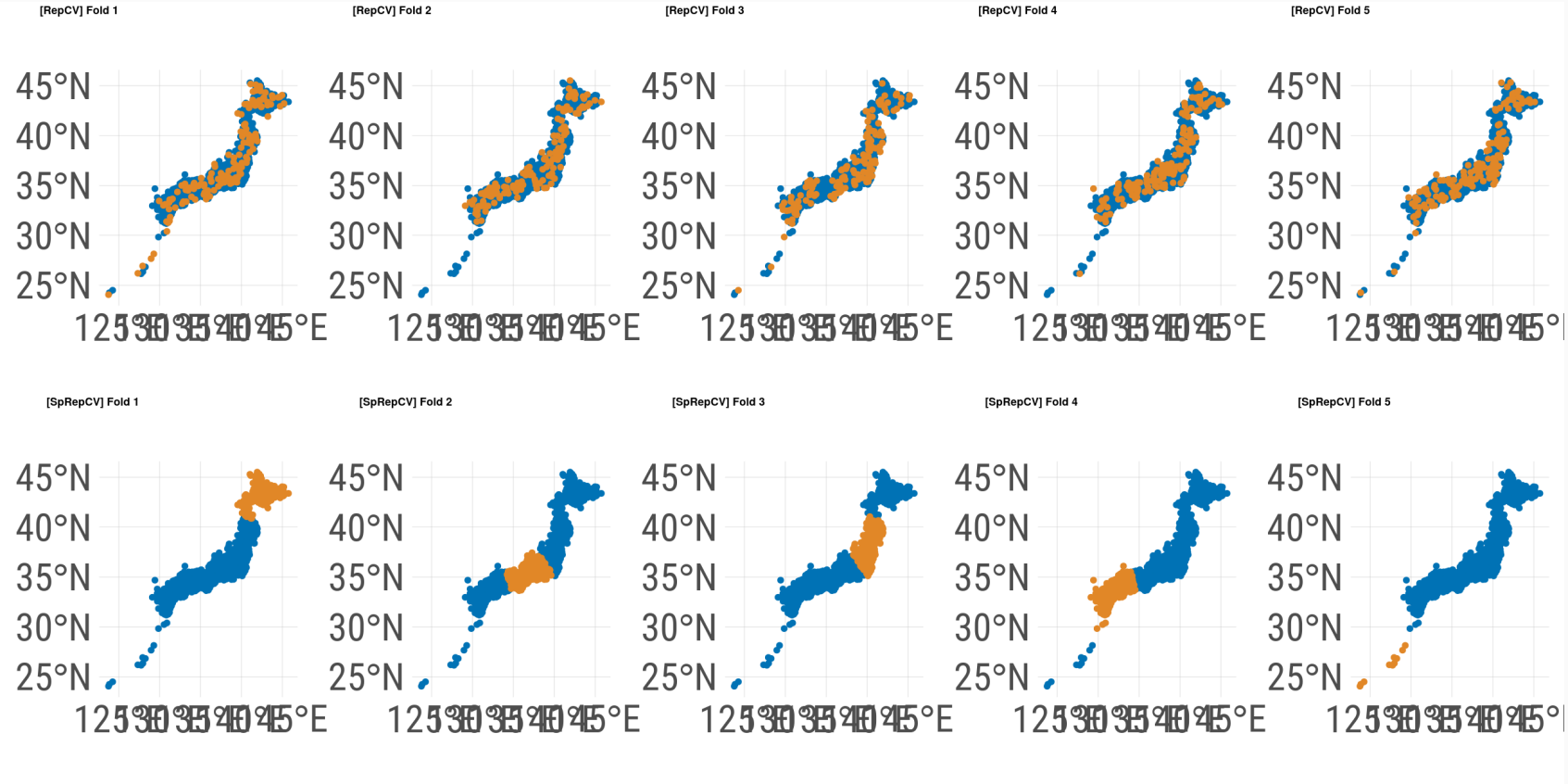
Spatial CV

```
resampling_sp <-
  makeResampleDesc("SpRepCV",
                   fold = 5, reps = 5)

set.seed(123)
sp_cv_out <-
  resample(learner    = learner.rf,
           task       = spatial.task,
           resampling = resampling_sp,
           measures   = list(auc))
mean(sp_cv_out$measures.test$auc, na.rm = TRUE)
# [1] 0.7624839
```

プロット

上段: k-fold CV samples 下段: Spatial CV samples



2.2 Target-oriented cross-validation

Target-oriented cross-validation

ざっくりいうと

- 空間 + 時間データの自己相関にも対応可能なCV
- サンプリングのデータの配置戦略を考慮する
    - LLO-CV… 特定のLocationデータのみをテストに
    - LTO-CV… 特定のTimeデータのみをテストに
    - LLTO-CV… 特定のLocation, Timeをテストに

CAST package

```
library(CAST)
library(caret)
```

Step

1. `CreateSpacetimeFolds()`で
   割り当てデータのインデックスを操作
2. `trainControl(index = )`に 生成したインデックスを指定

データの用意

```
df_train <-
  df_weather_20180815 %>%
  dplyr::select(station_no,
                elevation,
                temperature_mean,
                precipitation_sum) %>%
  st_set_geometry(NULL) %>%
  as.data.frame()
```

caretでtrain()

```
set.seed(123)
model <-
  caret::train(df_train[, c("elevation", "temperature_
        df_train$precipitation_sum,
      method     = "rf",
      tuneLength = 1,
      importance = TRUE,
      trControl  = trainControl(method = "cv",
                                number = 5))
```

caret で cv

```
model$results
```

```
##   mtry      RMSE  Rsquared       MAE  RMSESD RsquaredS
## 1    1 24.91898 0.2395822 15.06634 1.06227 0.047478
```

Target-oriented CV

考慮すべき変数（空間、時間）を
CreateSpacetimeFolds(spacevar =) に指定

```
indices <-
  CAST::CreateSpacetimeFolds(
    df_train,
    spacevar = "station_no",
    k        = 5,
    seed     = 123)
```

Target-oriented CV

```r
set.seed(123)

model_LLO <-
  caret::train(
    df_train[, c("elevation", "temperature_mean")],
    df_train$precipitation_sum,
    method     = "rf",
    tuneLength = 1,
    importance = TRUE,
    trControl  = trainControl(method = "cv",
                              index = indices$index))
```

Target-oriented CV

LLOの方がConventional CVよりもRMSEが低くなった

```
mean(model$results$RMSE)
```

```
## [1] 24.91898
```

```
mean(model_LLO$results$RMSE)
```

```
## [1] 24.50013
```

References

- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Arroita, G. G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929.
- Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental Modelling & Software, 101, 1–9.
- The importance of spatial cross-validation in predictive modeling
- Visualization of spatial cross-validation partitioning