Remarks on "Prediction of Regional Goods Demand Incorporating the Effect of Weather"

Watanabe Takuya, Hiroaki Muroi, Motoki Naruke Edirium K.K.
Kyoto Yono Nowcast, Inc.
Gen Kobayashi University of Tsukuba
Masanori Yamasaki Acroquest Technology Co., Ltd.

Presented at IEEE Workshop on Big Data Analytics in Manufacturing and Supply Chains,
2016 IEEE International Conference on Big Data.
Remarks for Nowcast Financial Big Data Seminar, 24 Jan. 2018.

Outline

- Motivation
- Data
  - daily sales records
  - daily weather records
- Models
  - sales prediction models
- Methods
  - three methods of prediction
- Results
  - predictive performance
  - interaction between item and area
  - interaction between area and weather
- Discussion
- Summary

Motivation

- Precise demand prediction of consumer goods is beneficial
  - supply chains can be optimised according to predicted demand
    - thus leading to higher profit
  - each item has different price elasticity
    - that is, effect of price change for sales is different for each item
  - goods demand is different in each area
  - effect of weather for goods demand would also be different in each area
- Big data can be utilised in two different ways
  - build complex models which need a vast amount of training data
    - image classification, speech recognition, machine translation etc.
  - sub-divide data according to some criteria for fine grained analysis
    - area, age, product model etc.
    - analyse each sub-divided data sets independently
    - being able to keep model's performance well because sub-divided data sets still have enough records

Data: Daily Sales Records

- Collected from Point of Sale (POS) systems
  - mainly of supermarkets in Japan
  - provided by Nikkei Digital Media, Inc.
  - 3 billion records, 1500 stores, 3 million items
  - from 1 Nov. 2011 to 31 Oct. 2015

Courtesy of Tsutomu Watanabe lab., the University of Tokyo

Dataset contains records from 1 Oct. 2009, but we exclude records of first two years to avoid effects of the earthquake in Mar. 2011.

- Consists of
  - store code
  - date
  - item code (Japan Article Number)
  - sales
  - quantity sold
    - target of the prediction
- Other tables
  - store code to area
  - item code to category

Sales Record Examples

Table I
RECORD EXAMPLES IN THE SALES TABLE. ITEM CODES ARE ANONYMISED.

Store Code	Date	Item Code	Sales	Quantity Sold
1	2009/10/01	**********713	3048	1
1	2009/10/01	**********132	1029	1
1	2009/10/01	**********550	760	2
1	2009/10/01	**********635	872	4
1	2009/10/01	**********026	284	1
...	...	...	...	...
2	2009/10/01	**********550	1140	3
2	2009/10/01	**********635	436	2
2	2009/10/01	**********924	3048	1
2	2009/10/01	**********940	2858	1
2	2009/10/01	**********430	2858	1
...	...	...	...	...

In JPY

Amount of Sales Records and Stores

Table III
TOP 5 AREAS IN THE SALES TABLE IN TERMS OF THE NUMBER OF STORES.

Prefecture	Number of Stores
Kanagawa	182
Hokkaido	116
Fukuoka	99
Tokyo	90
Osaka	68

Table II
TOP 10 ITEMS IN THE SALES TABLE.

Item Code	Number of Records
4902102072618	422693
4902705104167	422558
4902102069359	412273
45019517	412230
4902102084178	411119
4901411011523	409652
49722741	407743
4901411011547	406670
4903110063278	406483
4901340689213	404653

Data: Daily Weather Records

- Collected from the web site of Japan Meteorological Agency
  - http://www.jma.go.jp/jma/indexe.html
- Consists of 16 variables
  - atmospheric pressure at observatory/sea level
  - total/1h max/10m max precipitation
  - min/max/mean temperature etc.

Weather Record Examples

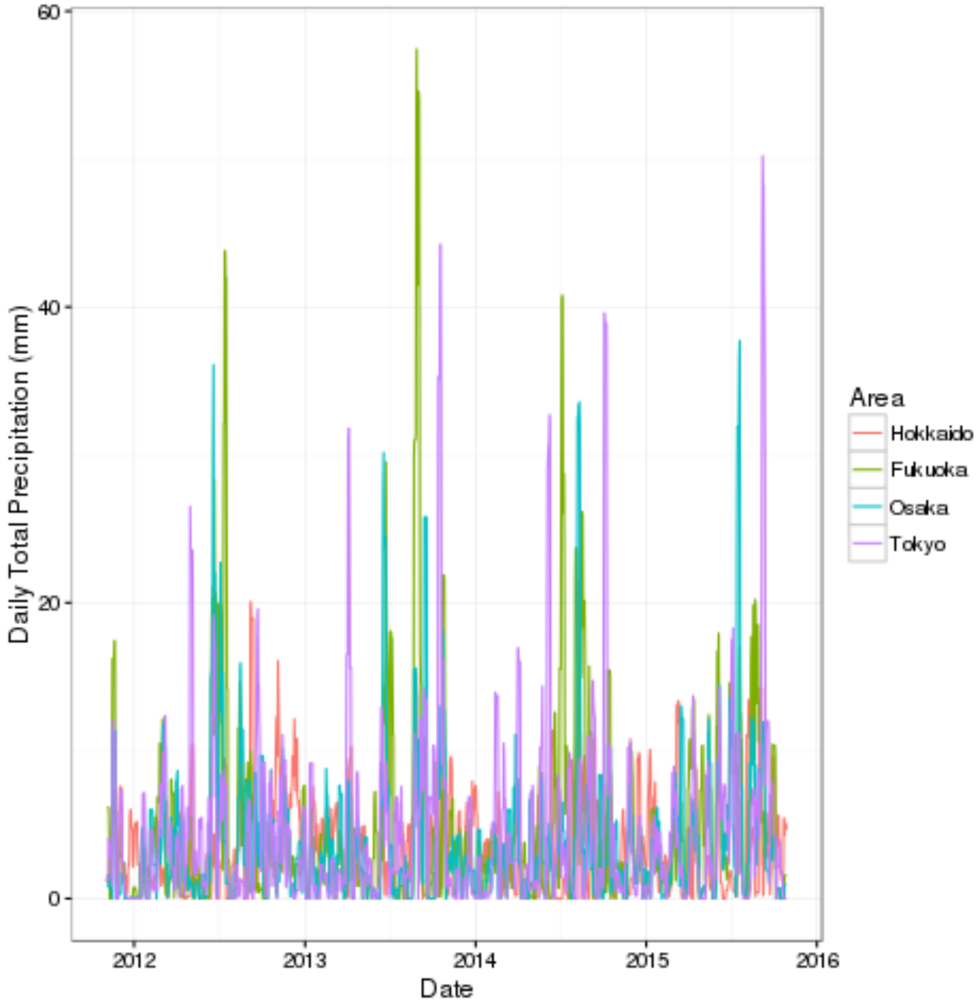


Figure 1. Record of daily total precipitation.

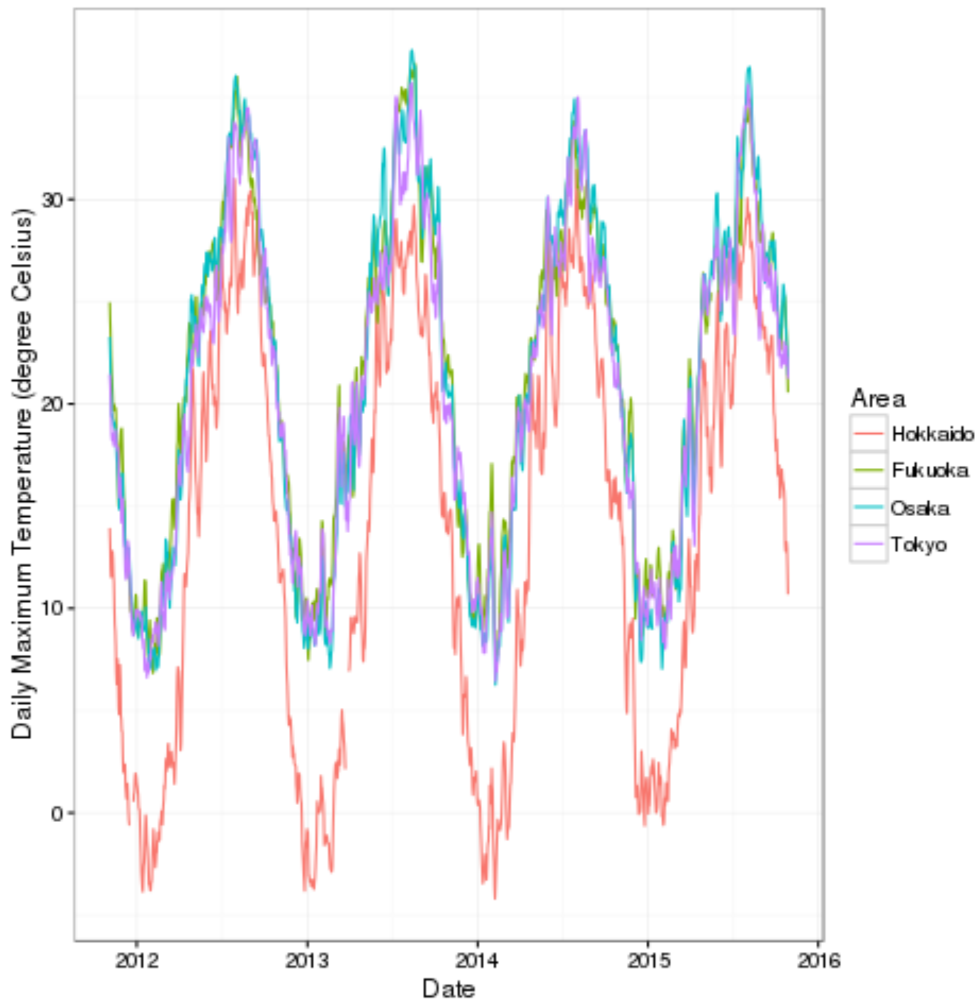


Figure 2. Record of daily maximum temperature.

Correlation Matrix of Weather Metrics

Table IV
CORRELATION MATRIX OF WEATHER METRICS.

Pressure at Observatory	1	0.49	-0.09	-0.11	-0.11	-0.14	-0.16	-0.13	-0.23	-0.13	0.04	-0.07	-0.10	0.08	-0.06	0.00
Pressure at Sea Level	0.49	1	-0.26	-0.27	-0.26	-0.48	-0.46	-0.50	-0.32	-0.29	-0.16	-0.21	-0.24	0.15	-0.00	0.00
Total Precipitation	-0.09	-0.26	1	0.85	0.73	0.08	0.04	0.12	0.41	0.39	0.09	0.15	0.17	-0.37	0.07	-0.00
1h Max Precipitation	-0.11	-0.27	0.85	1	0.92	0.14	0.10	0.17	0.38	0.36	0.07	0.14	0.17	-0.34	0.03	-0.02
10m Max Precipitation	-0.11	-0.26	0.73	0.92	1	0.16	0.13	0.20	0.38	0.35	0.05	0.13	0.16	-0.33	0.02	-0.01
Mean Temperature	-0.14	-0.48	0.08	0.14	0.16	1	0.99	0.99	0.28	0.29	-0.08	-0.13	-0.17	0.15	-0.24	-0.32
Max Temperature	-0.16	-0.46	0.04	0.10	0.13	0.99	1	0.95	0.21	0.18	-0.10	-0.13	-0.17	0.24	-0.25	-0.33
Min Temperature	-0.13	-0.50	0.12	0.17	0.20	0.99	0.95	1	0.35	0.38	-0.07	-0.12	-0.17	0.04	-0.22	-0.30
Mean Humidity	-0.23	-0.32	0.41	0.38	0.38	0.28	0.21	0.35	1	0.90	-0.18	-0.14	-0.17	-0.56	0.10	0.03
Min Humidity	-0.13	-0.29	0.39	0.36	0.35	0.29	0.18	0.38	0.90	1	-0.11	-0.13	-0.16	-0.61	0.05	0.01
Mean Wind Speed	0.04	-0.16	0.09	0.07	0.05	-0.08	-0.10	-0.07	-0.18	-0.11	1	0.85	0.80	0.00	0.05	0.05
Max Wind Speed	-0.07	-0.21	0.15	0.14	0.13	-0.13	-0.13	-0.12	-0.14	-0.13	0.85	1	0.93	-0.01	0.09	0.08
Instant. Wind Speed	-0.10	-0.24	0.17	0.17	0.16	-0.17	-0.17	-0.17	-0.17	-0.16	0.80	0.93	1	-0.04	0.10	0.11
Sunlight Hours	0.08	0.15	-0.37	-0.34	-0.33	0.15	0.24	0.04	-0.56	-0.61	0.00	-0.01	-0.04	1	-0.12	-0.08
Snowfall	-0.06	-0.00	0.07	0.03	0.02	-0.24	-0.25	-0.22	0.10	0.05	0.05	0.09	0.10	-0.12	1	0.46
Snow Depth	0.00	0.00	-0.00	-0.02	-0.01	-0.32	-0.33	-0.30	0.03	0.01	0.05	0.08	0.11	-0.08	0.46	1

Weather Variable Selection

- Discard highly correlated variables
- Selected variables
  - average atmospheric pressure at observatory
  - total precipitation
  - maximum temperature
  - minimum humidity
  - maximum wind speed
  - sunlight hours
  - snow depth

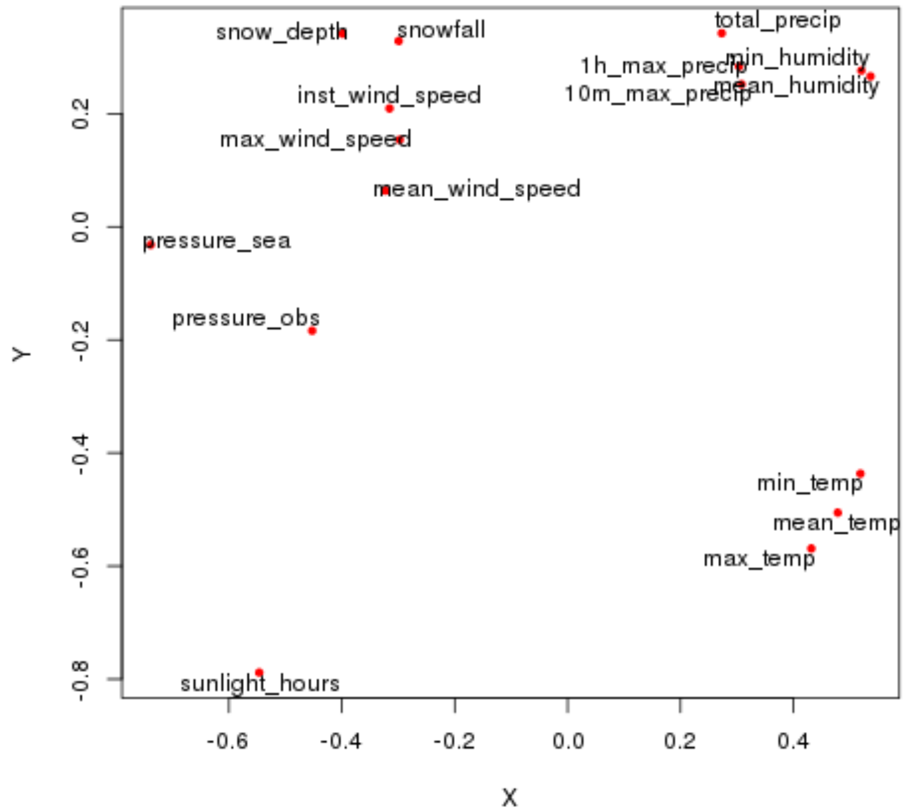


Figure 3. MDS plot of the relationships between weather metrics. Axes have no particular meanings.

Relation between Price and Quantity Sold

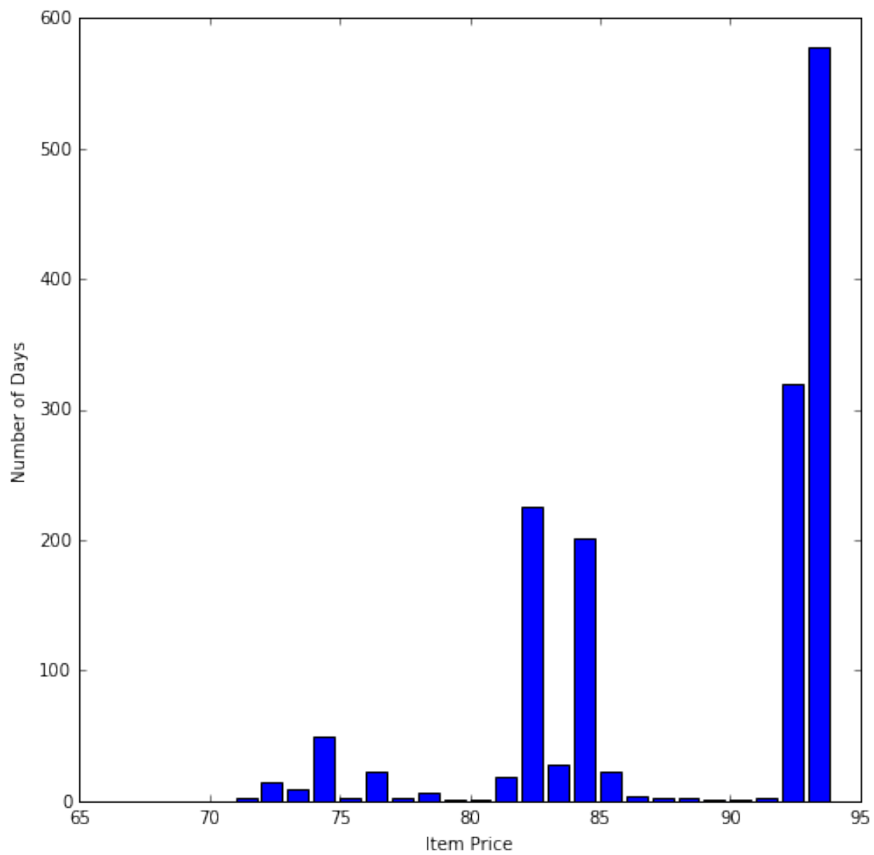


Figure 4. Histogram of price showing the number of days in which an item is sold at each price bin.

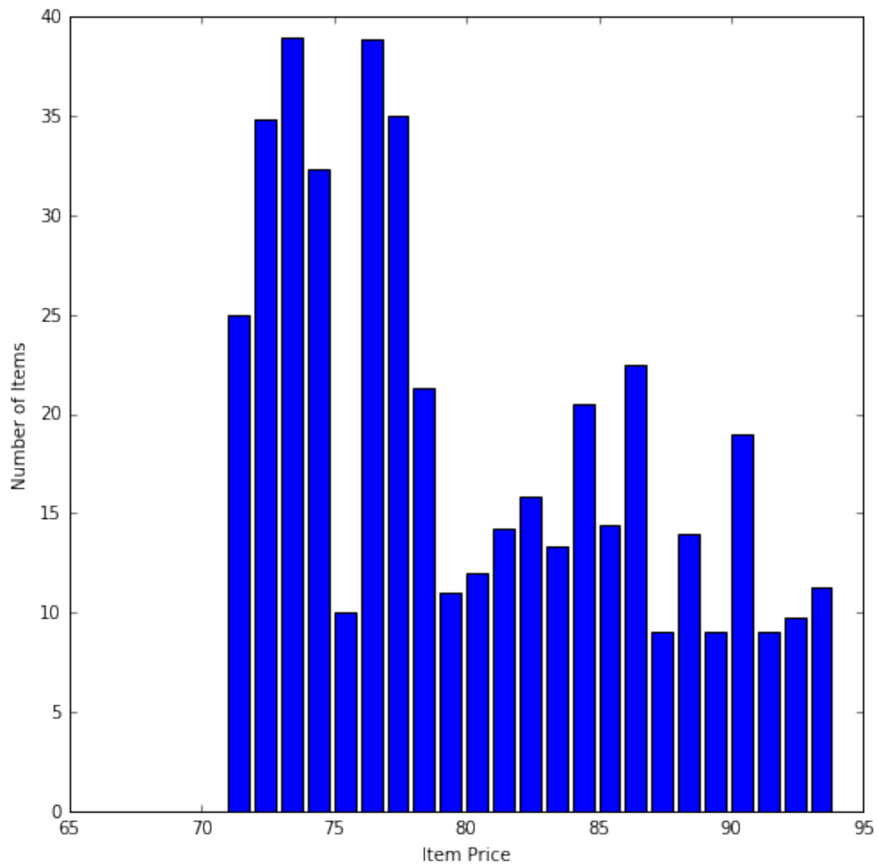


Figure 5. Average daily quantity sold per price bin.

Mode Price and Sale Flag

- There would be an effect of "sale" itself
  - sales change which could not be explained by price change
- Calculate mode prices for each day
  - from the prices of the period of 28 days before and after each day
- Sale flag condition
  - prices below 1σ of the mode

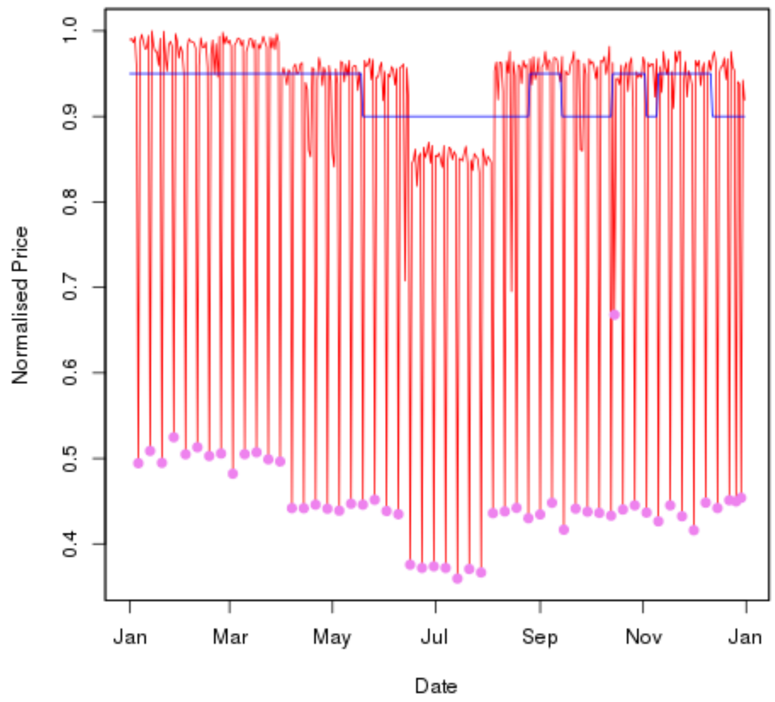


Figure 6. Price record example with mode price and sale flags. Red and blue lines represent price and mode price respectively, and magenta dots indicate days where the sale flag is on.

Prediction Models

- Explanatory variables
  - quantity sold in the previous day
  - day of the week
  - season
  - weather related variables
  - price
  - sale flag

Japanese holidays are also taken into account as a separate variable.

- Target variable: quantity sold
- Many fine grained models
  - dataset is separated by item and area
- Three prediction methods
  - linear regression
  - neural network
  - deep neural network

Method: Linear Regression

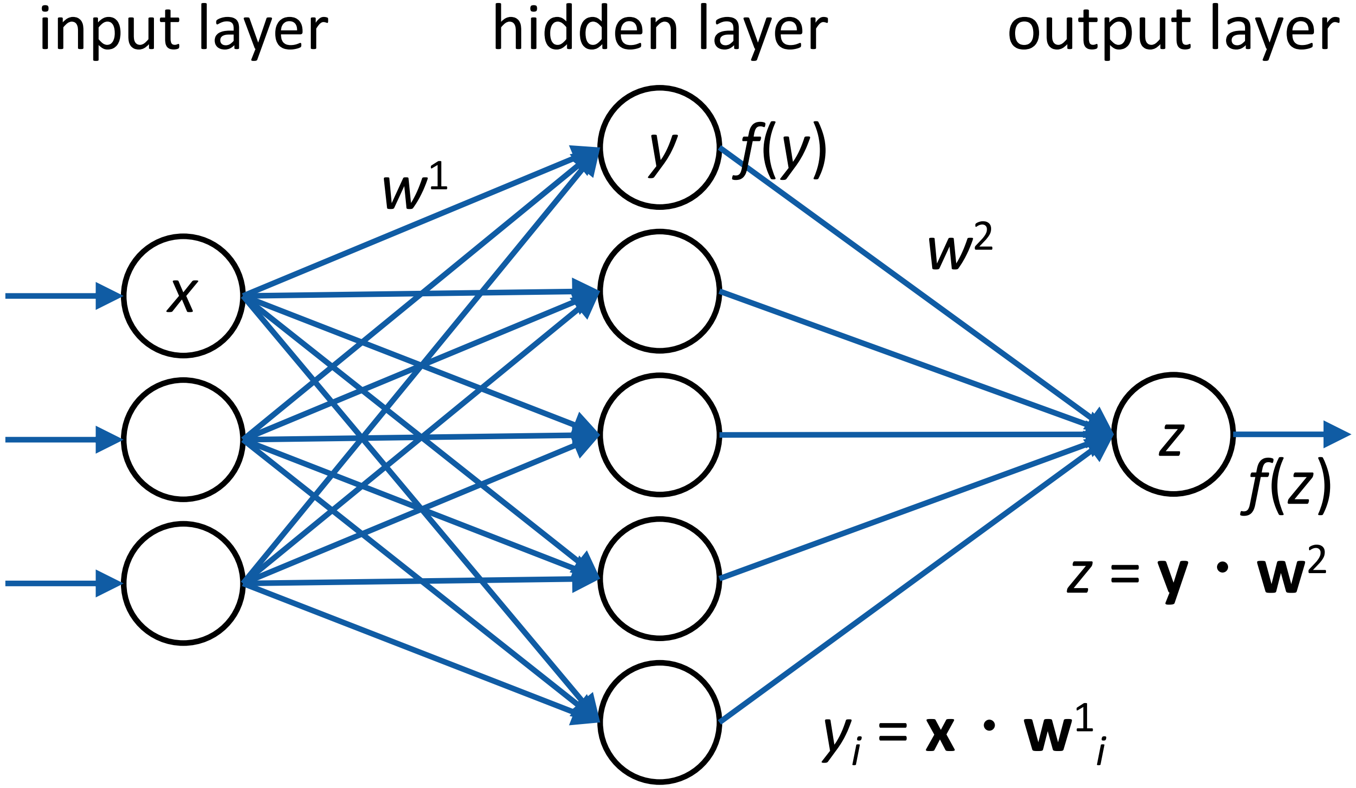
- Simple linear regression
  - not much to say

Categorical variables are converted to dummy variables. As a result, the number of explanatory variables are 20 in total.
The model is fitted without regularization.

Method: Neural Network

- Basics
  - weight for each connection, activation function for each node
- Simple two layer network
  - only one hidden layer
  - single output node for regression

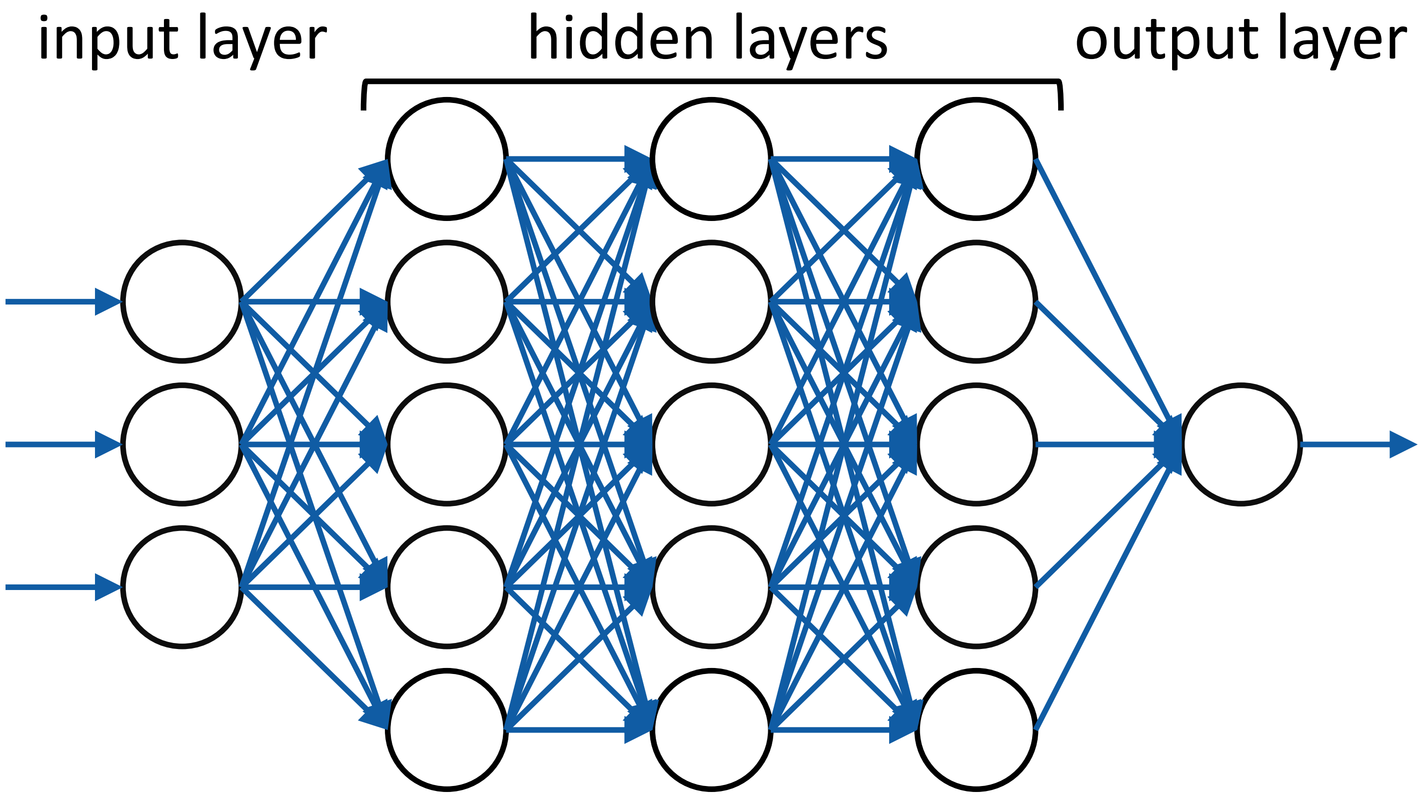
3 nodes in the hidden layer in our model.



Method: Deep Neural Network

- Deeper structure
  - several number of hidden layers
    - actually not so deep, but adequate for this kind of simple regression problems?
  - single output node for regression (not changed)

3 hidden layers and 10 nodes in each hidden layer in our model.

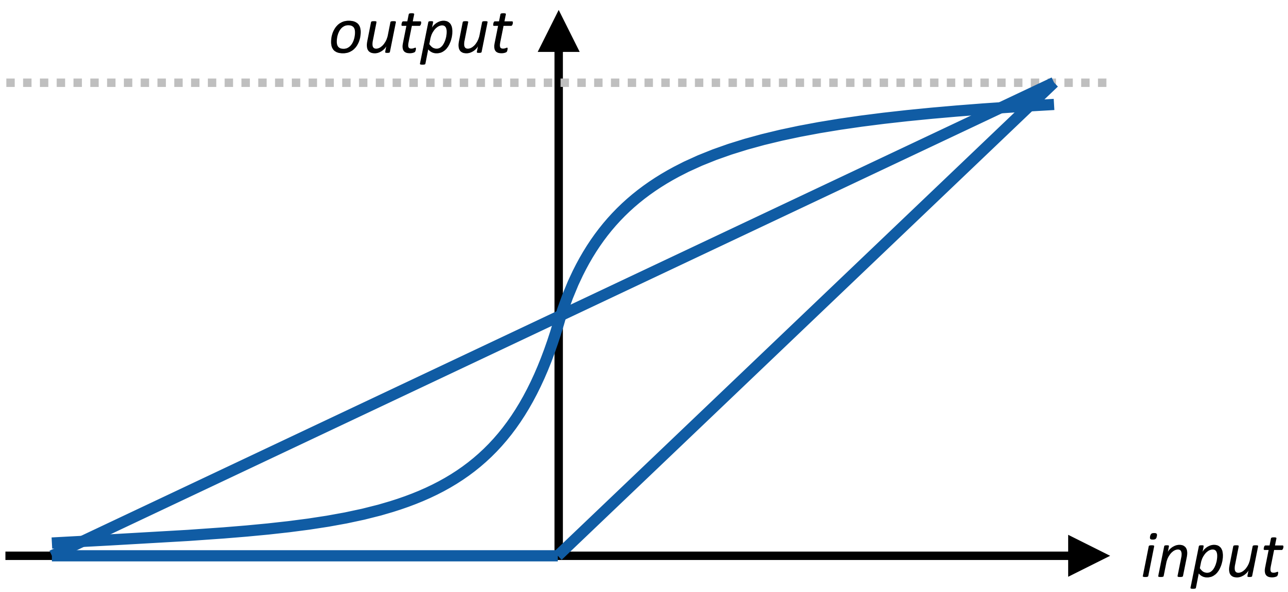


Activation Functions

Effectively the same as PCA.

- Linear activation function (no activation function) for our NN
- Sigmoid function, traditionally used for classification tasks
- Rectified linear unit for our DNN
  - effective for faster convergence

Linear for the output layer.



System for Analysis

- Language
  - R and Python
- Libraries
  - *caret* for hyper parameter tuning
  - *nnet* for NN, *h2o* for DNN
- Database
  - PostgreSQL
    - single node database built on an Azure instance
    - approx. 1 TB data size including indices
    - too slow, even for relatively simple queries
  - Impala
    - rebuilt the database to seek higher performance
    - single node database built on an AWS EC2 instance
    - utilised I2 instance's SSD based instance storage for Avro format tables
    - significantly faster

Results: Performance

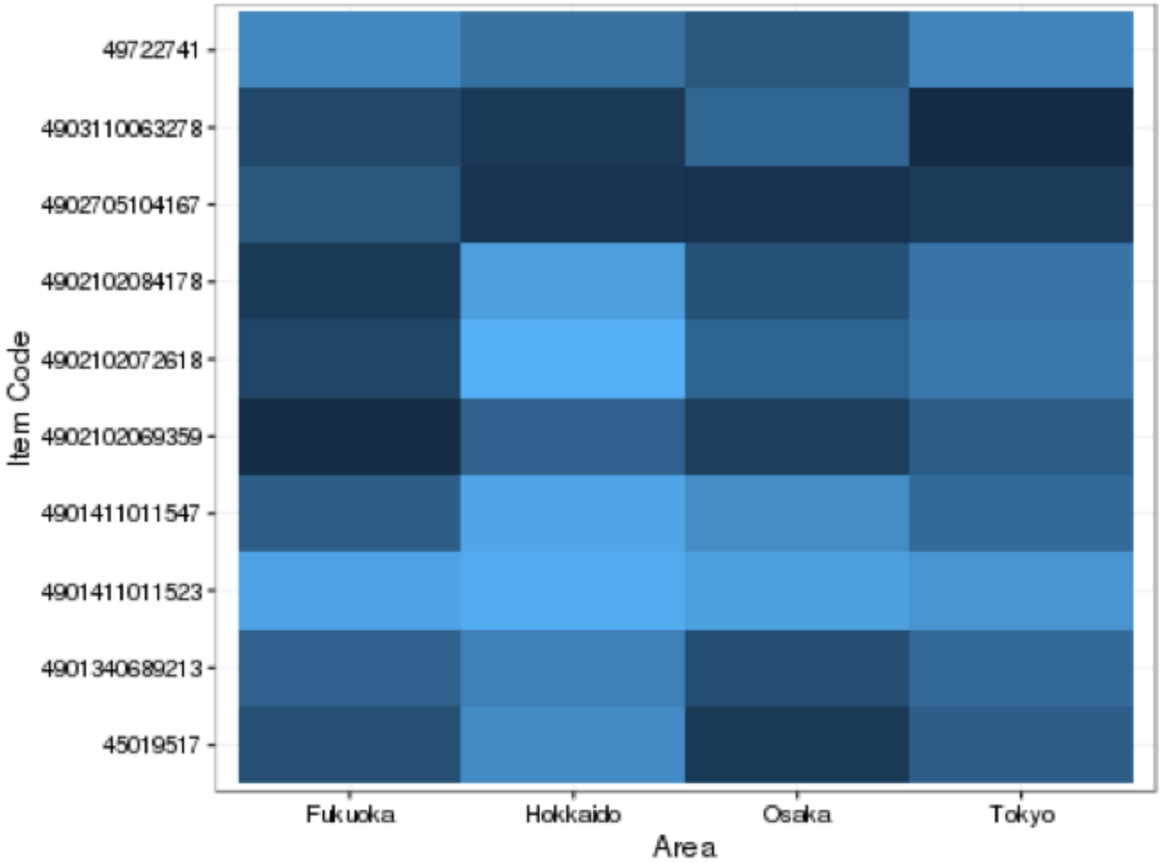
To absorb the sales level difference of each store.

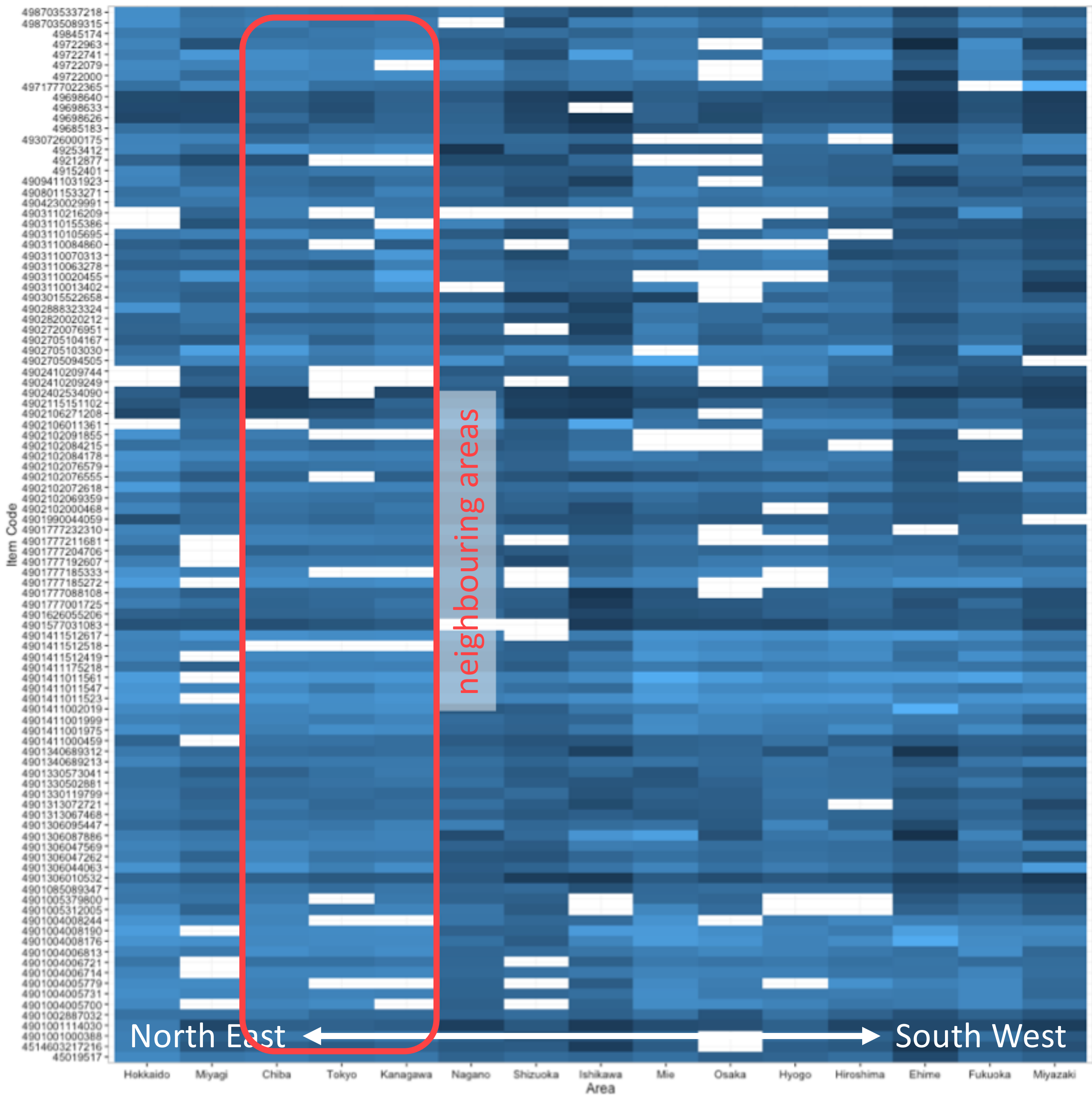
- Quantity sold is normalised to [0, 1] interval
  - predictive performance can simply be compared by root mean squared error (RMSE)
- Measured RMSE with 5-fold cross validation
- Neural network seems better than linear regression
- Deep neural network may be better than neural network

Area	RMSE (Linear)	RMSE SD (Linear)	RMSE (Neural)	RMSE SD (Neural)	RMSE (Deep)	RMSE SD (Deep)
Fukuoka	0.11320	0.00089	0.11137	0.00170	<u>0.11016</u>	0.00104
Hokkaido	0.09364	0.00356	0.09259	0.00327	<u>0.08527</u>	0.00101
Osaka	0.09448	0.00151	0.08918	0.00174	<u>0.08801</u>	0.00098
Tokyo	0.10308	0.00095	<u>0.09878</u>	0.00296	0.09934	0.00160

Results: Interactions between Item and Area

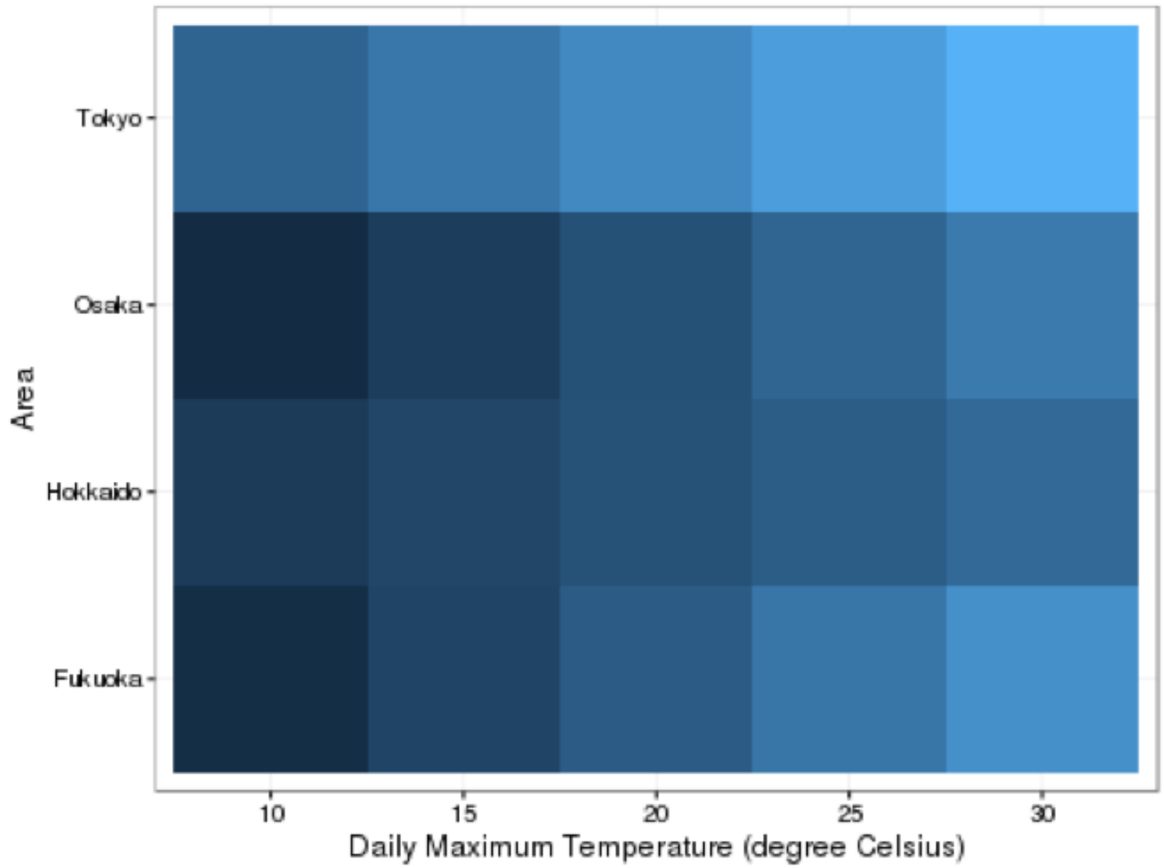
- Obtain demand curve for each item by changing the price
- Visualise demand by integrating the demand curve
  - lighter colour indicates more total demand
  - each item exhibits different pattern




neighbouring areas
North East
South West
Item Code
Hokkaido
Miyagi
Chiba
Tokyo
Kanagawa
Nagano
Shizuoka
Ishikawa
Mie
Osaka
Hyogo
Hiroshima
Ehime
Fukuoka
Miyazaki
Area

Results: Interactions between Area and Weather

- Visualise demand by a similar way as the item x area case
- Example: an ice cream product
  - higher temperature draws more demand
  - above effect is not apparent in Hokkaido (northern island)



Discussions

- Validity of the models and observations
  - a single big model with area variables could possibly perform better
  - no statistical check is done as to whether there is a real difference
- Effect of mid-term changes of economical circumstances
  - would be better to incorporate yearly trends into our models
- Effect of associations between items
  - would be better to incorporate category/manufacturer information into our models
- Better (meta-)models and/or prediction methods
  - there might be more suitable neural network models for this kind of regression problems
  - other prediction methods could be applied

Reimplementation by using Keras/TensorFlow is under way.

Random forest didn't perform well.

Summary

- Analysed daily sales records (POS data) and weather records
  - to predict goods demand
- Compared performance of three prediction method
  - linear regression < neural network < deep neural network...?
- Visualised total demand by integrating the demand curve
  - interactions between item and area
  - interactions between area and weather

Thank you!

These slides will soon be available at our web site:
http://www.edirium.co.jp/