

Parametrized event records data transformations

Anton Antonov
MathematicaForPrediction at WordPress
MathematicaForPrediction at GitHub
MathematicaVsR at GitHub
August-October 2018

Version 1.0

Introduction

In this document we describe transformations of events records data in order to make that data more amenable for the application of Machine Learning (ML) algorithms. Consider the following **problem formulation** that is done with the next five bullet points.

- From data representing a (most likely very) diverse set of events we want to derive contingency matrices corresponding to each of the variables in that data.
- The events are observations of the values of a certain set of variables for a certain set of entities. Not all entities have events for all variables.
- The observation times do not form a regular time grid.
- Each contingency matrix has rows corresponding to the entities in the data and has columns corresponding to time.
- The software component providing the functionality should allow parametrization and repeated execution. (As in ML classifier training and testing scenarios.)

The phrase “event records data” is used instead of “time series” in order to emphasize that (i) some variables have categorical values, and (ii) the data can be given in some general database form, like transactions long-form.

The required transformations of the event records in the problem formulation above are done through the monad ERTMon, [AAp3]. (The name ERTMon comes from “**E**vent **R**ecords **T**ransformations **M**onad”.)

The monad code generation and utilization is explained in [AA1] and implemented with [AAp1].

It is assumed that the event records data is put in a form that makes it (relatively) easy to extract time series for the set of entity-variable pairs present in that data.

In brief ERTMon performs the following sequence of transformations.

1. The event records of each entity-variable pair are shifted to adhere to a specified start or end point,
2. The event records for each entity-variable pair are aggregated and normalized with specified functions over a specified regular grid,
3. Entity vs. time interval contingency matrices are made for each combination of variable and aggregation function.

The transformations are specified with a “computation specification” dataset.

Here is an example of an ERTMon pipeline over event records:

Out[*n*]=

ERTMonUnit[] =>
ERTMonSetEventRecords[eventRecords] =>
ERTMonSetEntityAttributes[entityAttributes] =>
ERTMonSetComputationSpecification[compSpec] =>
ERTMonGroupEntityVariableRecords =>
ERTMonEntityVariableGroupsToTimeSeries["MaxTime"] =>
ERTMonAggregateTimeSeries

initialize the monad pipeline
ingest event records
ingest event attributes
ingest computation specification
group records per {EntityID, Variable}
align the event records and create time series
aggregate the time series
(according to the computation specification)

The rest of the document describes in detail:

- the structure, format, and interpretation of the event records data and computations specifications,
- the transformations of time series aligning, aggregation, and normalization,
- the software pattern design -- a monad -- that allows sequential specifications of desired transformations.

Concrete examples are given using weather data. See [AAp9].

Package load

The following commands load the packages [AAp1-AAp9].

```
In[108]:= Import["https://raw.githubusercontent.com/antononcube/MathematicaForPrediction/master/MonadicProgramming/MonadicEventRecordsTransformations.m"]
Import["https://raw.githubusercontent.com/antononcube/MathematicaForPrediction/master/MonadicProgramming/MonadicTracing.m"]
Import["https://raw.githubusercontent.com/antononcube/MathematicaForPrediction/master/Misc/WeatherEventRecords.m"]
```

Data load

The data we use is weather data from meteorological stations close to certain major cities. We retrieve the data with the function WeatherEventRecords from the package [AAp9].

```
In[111]:= ? WeatherEventRecords

WeatherEventRecords[ citiesSpec_: {{_String, _String}..}, dateRange: {{_Integer, _Integer, _Integer}, {_Integer, _Integer, _Integer}}, wProps: {_String..}: {"Temperature"}, nStations_Integer: 1] gives an association with event records data.
```

```
In[112]:= citiesSpec = {"Miami", "USA"}, {"Chicago", "USA"}, {"London", "UK"};
dateRange = {{2017, 7, 1}, {2018, 6, 31}};
wProps = {"Temperature", "MaxTemperature", "Pressure", "Humidity", "WindSpeed"};
res1 = WeatherEventRecords[citiesSpec, dateRange, wProps, 1];

In[116]:= citiesSpec = {"Jacksonville", "USA"}, {"Peoria", "USA"}, {"Melbourne", "Australia"};
dateRange = {{2016, 12, 1}, {2017, 12, 31}};
res2 = WeatherEventRecords[citiesSpec, dateRange, wProps, 1];
```

Here we assign the obtained datasets to variables we use below:

```
In[119]:= eventRecords = Join[res1["eventRecords"], res2["eventRecords"]];
entityAttributes = Join[res1["entityAttributes"], res2["entityAttributes"]];
```

Here are the summaries of the datasets eventRecords and entityAttributes:

```
In[121]:= RecordsSummary[eventRecords]
```

1 EntityID	2 LocationID	3 ObservationTime	4 Variable	5 Value
KNIP	1967 Jacksonville	1967 Min	3 689 539 200	Min -19.33
WMO95866	1965 Melbourne	1965 1st Qu	3 705 955 200	1st Qu 2.78
KMDW	1802 , Chicago	1802 , Mean	3.71474×10^9	Median 16.11
KMFL	1798 Miami	1798 Median	3 715 113 600	3rd Qu 26.78
KGEU	1572 Peoria	1572 3rd Qu	3 723 235 200	Mean 153.606
EGLC	1456 London	1456 Max	3 739 392 000	Max 1043.1

In[122]:= RecordsSummary[entityAttributes]

1 EntityID		3 Value	
EGLC	2	USA	4
KGEU	2	Australia	1
KMDW	2	City	6
		Chicago	1
KMFL	2	Country	6
		Jacksonville	1
KNIP	2	London	1
WM095866	2	Melbourne	1
		(Other)	3

Out[122]= {

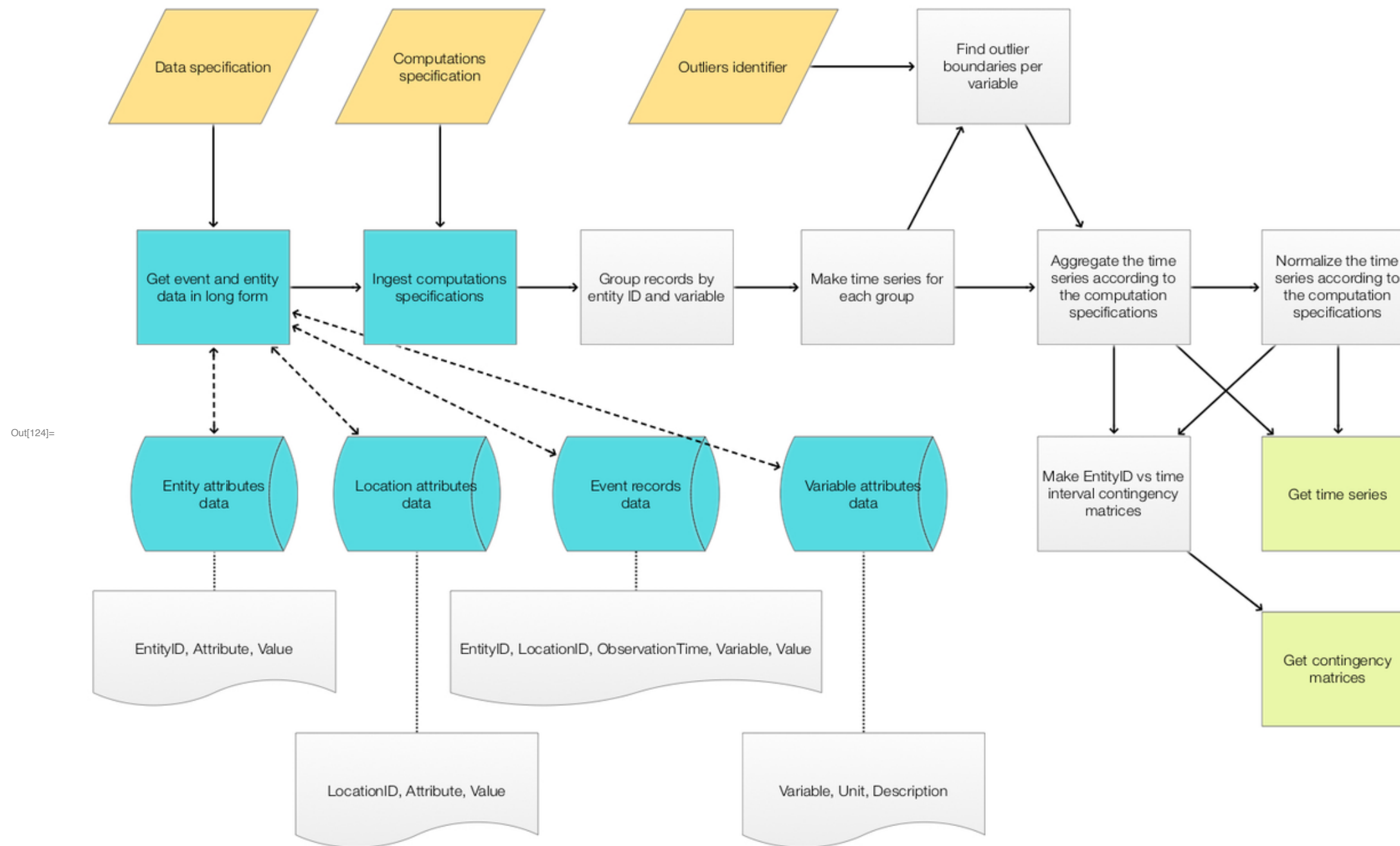
Design considerations

Workflow

The steps of the main event records workflow addressed in this document follow.

1. Ingest event records and entity attributes given in the Star schema style.
2. Ingest a computation specification.
 - 2.1. Specified are aggregation time intervals, aggregation functions, normalization types and functions.
3. Group event records based on unique entity ID and variable pairs.
 - 3.1. Additional filtering can be applied using the entity attributes.
4. For each variable find descriptive statistics properties.
 - 4.1. This is to facilitate normalization procedures.
 - 4.2. Optionally, for each variable find outlier boundaries.
5. Align each group of records to start or finish at some specified point.
 - 5.1. For each variable we want to impose a regular time grid.
6. From each group of records produce a time series.
7. For each time series do prescribed aggregation and normalization.
 - 7.1. The variable that corresponds to each group of records has at least one (possibly several) computation specifications.
8. Make a contingency matrix for each time series obtained in the previous step.
 - 8.1. The contingency matrices have entity ID's as rows, and time intervals enumerating values of time intervals.

The following flow-chart corresponds to the list of steps above.



A corresponding monadic pipeline is given in the section “Larger example pipeline”.

Feature engineering perspective

The workflow above describes a way to do feature engineering over a collection of event records data. For a given entity ID and a variable we derive several different time series. Couple of examples follow.

- One possible derived feature (times series) is for each entity-variable pair we make time series of the hourly mean value in each of the eight most recent hours for that entity. The mean values are normalized by the average values of the records corresponding to that entity-variable pair.
- Another possible derived feature (time series) is for each entity-variable pair to make a time series with the number of outliers in the each half-hour interval, considering the most recent 20 half-hour intervals. The outliers are found by using outlier boundaries derived by analyzing all values of the corresponding variable, across all entities.

From the examples above -- and some others -- we conclude that for each feature we want to be able to specify:

- aggregation interval length,
- aggregation function (to be applied in each interval),
- maximum history length (say from the most recent observation),
- normalization function (per entity, per cohort of entities, per variable),
- conversion of categorical values into numerical ones.

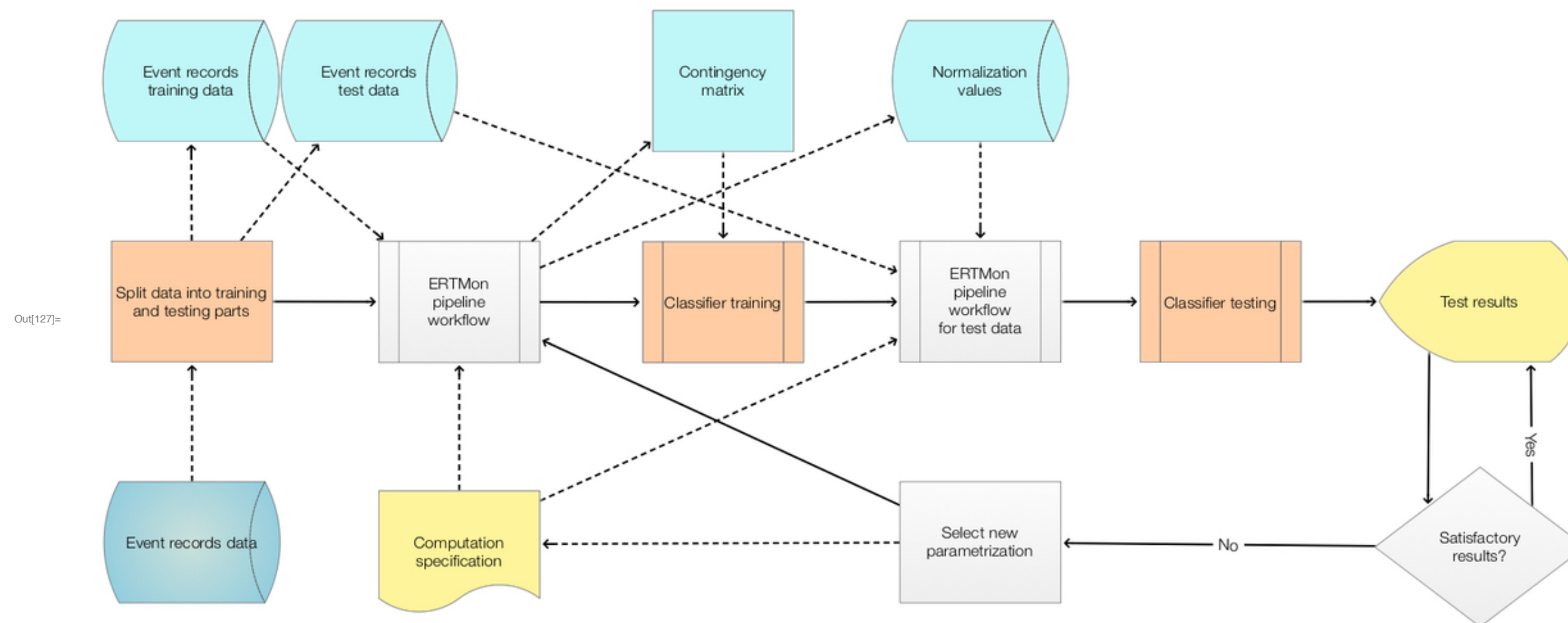
Repeated execution

We want to be able to do repeated executions of the specified workflow steps.

Consider the following scenario. After the event records data is converted to a entity-vs-feature contingency matrix, we use that matrix to train and test a classifier. We want to find the combination of features that gives the best classifier results. For that reason we want to be able to easily and systematically change the computation specifications (interval size, aggregation and normalization functions, etc.) With different computation specifications we obtain different entity-vs-feature contingency matrices, that would have different performance with different classifiers.

Using the classifier training and testing scenario we see that there is another repeated execution perspective: after the feature engineering is done over the training data, we want to be able to execute exactly the same steps over the test data. Note that with the training data we find certain global or cohort normalization values and outlier boundaries that have to be used over the test data. (Not derived from the test data.)

The following diagram further describes the repeated execution workflow.



Further discussion of making and using Machine Learning classification workflows through the monad software design pattern can be found in [AA2].

Event records data design

The data is structured to follow the style of Star schema. We have event records dataset (table) and entity attributes dataset (table).

The structure datasets (tables) proposed satisfy a wide range of modeling data requirements. (Medical and financial modeling included.)

Entity event data

The entity event data has the columns "EntityID", "LocationID", "ObservationTime", "Variable", "Value".

```
In[129]:= RandomSample[eventRecords, 6]
```

Out[129]=

EntityID	LocationID	ObservationTime	Variable	Value
KNIP	Jacksonville	3 690 403 200	Temperature	10.72
EGLC	London	3 721 593 600	Humidity	0.747
EGLC	London	3 724 876 800	Humidity	0.7
WMO95866	Melbourne	3 708 374 400	Humidity	0.787
KNIP	Jacksonville	3 707 856 000	WindSpeed	13.15
WMO95866	Melbourne	3 703 363 200	Temperature	10.83

Most events can be described through “Entity event data”. The entities can be anything that produces a set of event data: financial transactions, vital sign monitors, wind speed sensors, chemical concentrations sensors. The locations can be anything that gives the events certain "spatial" attributes: medical units in hospitals, sensors geo-locations, tiers of financial transactions.

Entity attributes data

The entity attributes dataset (table) has attributes (immutable properties) of the entities. (Like, gender and race for people, longitude and latitude for wind speed sensors.)

```
In[130]:= entityAttributes[[1 ;; 6]]
```

Out[130]=

EntityID	Attribute	Value
KMFL	City	Miami
KMFL	Country	USA
KMDW	City	Chicago
KMDW	Country	USA
EGLC	City	London
EGLC	Country	UK

Example

For example, here we take all weather stations in USA:

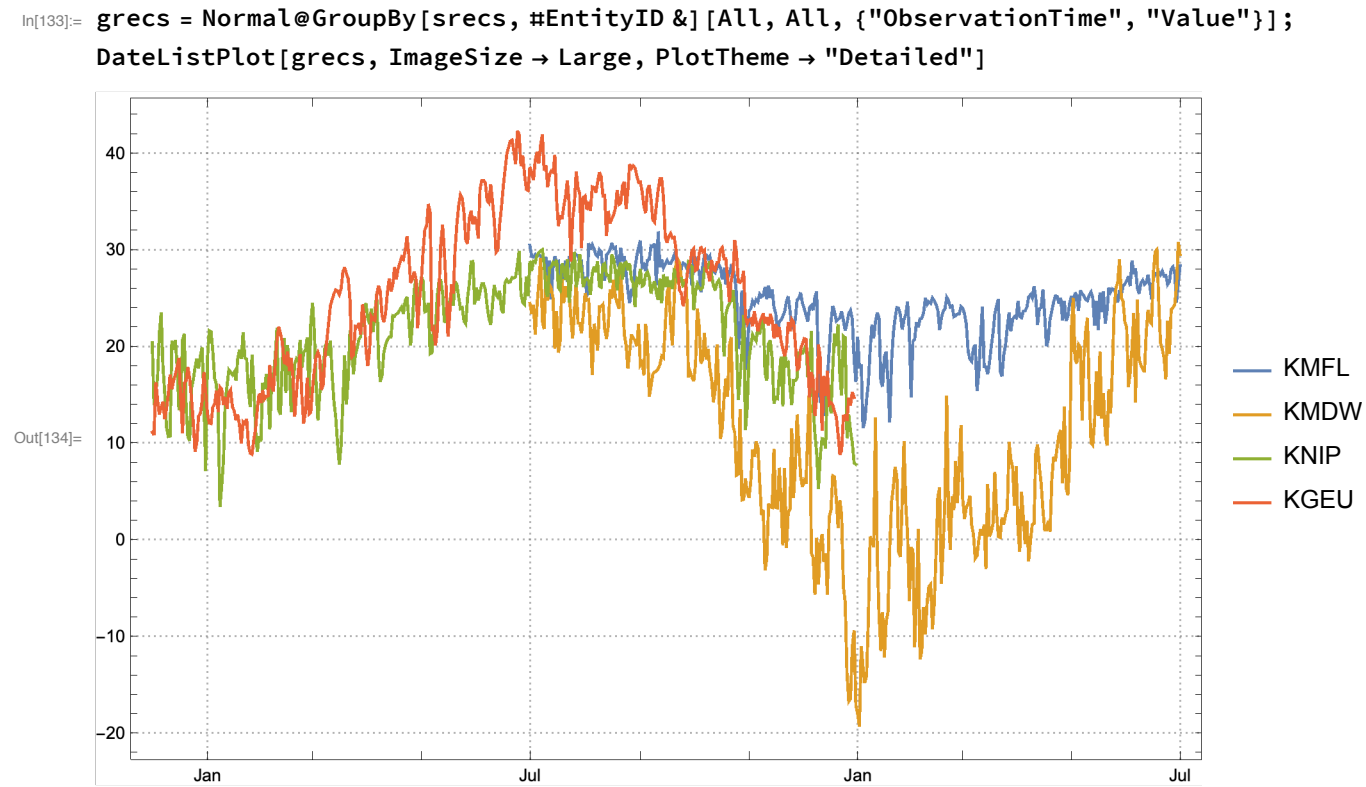
```
In[131]:= ws = Normal[entityAttributes[Select[#Attribute == "Country" && #Value == "USA" &], "EntityID"]]
```

Out[131]= { KMFL, KMDW, KNIP, KGEU }

Here we take all temperature event records for those weather stations:

```
In[132]:= srecs = eventRecords[Select[#Variable == "Temperature" && MemberQ[ws, #EntityID] &]];
```

And here plot the corresponding time series obtained by grouping the records by station (entity ID's) and taking the columns “ObservationTime” and ”Value”:



Monad elements

This section goes through the steps of the general ERTMon workflow.

Monad unit

The monad is initialized with ERTMonUnit.

```
In[189]:= ERTMonUnit[]
Out[189]= ERTMon[None, <| |>]
```

Ingesting event records and entity attributes

The event records dataset (table) and entity attributes dataset (table) are set with corresponding setter functions. Alternatively, they can be read from files in a specified directory.


```
In[190]:= p =
  ERTMonUnit[] =>
    ERTMonSetEventRecords[eventRecords] =>
      ERTMonSetEntityAttributes[entityAttributes] =>
        ERTMonEchoDataSummary;
```

» Data summary: $\langle \left| \text{eventRecords} \rightarrow \left\{ \begin{array}{llll} \text{EntityID} & \text{LocationID} & \text{ObservationTime} & \text{Variable} \end{array} \right. \right. \rangle$

1 EntityID	2 LocationID	3 ObservationTime	4 Variable	5 Value
KNIP	1967	Jacksonville	1967	Min 3 689 539 200
WM095866	1965	Melbourne	1965	1st Qu 3 705 955 200
KMDW	1802	Chicago	1802	Mean 3.71474×10^9
KMFL	1798	Miami	1798	Median 3 715 113 600
KGEU	1572	Peoria	1572	3rd Qu 3 723 235 200
EGLC	1456	London	1456	Max 3 739 392 000

\rangle , entityAttributes $\rightarrow \left\{ \begin{array}{llll} \text{EntityID} & \text{Attribute} & \text{Value} \end{array} \right. \rangle$

1 EntityID	2 Attribute	3 Value
EGLC	2	USA 4
KGEU	2	Australia 1
KMDW	2	Chicago 1
KMFL	2	Jacksonville 1
KNIP	2	London 1
WM095866	2	Melbourne 1
		(Other) 3

Computation specification

Using the package [Aap3] we can create computation specification dataset. Below is given an example of constructing a fairly complicated computation specification.

The package EmptyComputationSpecificationRow can be used to construct the rows of the specification.

```
In[191]:= EmptyComputationSpecificationRow[]
Out[191]:= <|Variable -> Missing[], Explanation -> , MaxHistoryLength -> 3600, AggregationIntervalLength -> 60, AggregationFunction -> Mean, NormalizationScope -> Entity, NormalizationFunction -> None|>

In[192]:= compSpecRows = Join[EmptyComputationSpecificationRow[], <|"Variable" -> #, "MaxHistoryLength" -> 60 * 24 * 3600, "AggregationIntervalLength" -> 2 * 24 * 3600,
  "AggregationFunction" -> "Mean", "NormalizationScope" -> "Entity", "NormalizationFunction" -> "Mean"|>] & /@ Union[Normal[eventRecords[All, "Variable"]]];
compSpecRows =
  Join[
    compSpecRows, Join[EmptyComputationSpecificationRow[], <|"Variable" -> #, "MaxHistoryLength" -> 60 * 24 * 3600, "AggregationIntervalLength" -> 2 * 24 * 3600,
      "AggregationFunction" -> "Range", "NormalizationScope" -> "Country", "NormalizationFunction" -> "Mean"|>] & /@ Union[Normal[eventRecords[All, "Variable"]]],
    Join[EmptyComputationSpecificationRow[], <|"Variable" -> #, "MaxHistoryLength" -> 60 * 24 * 3600, "AggregationIntervalLength" -> 2 * 24 * 3600,
      "AggregationFunction" -> "OutliersCount", "NormalizationScope" -> "Variable"|>] & /@ Union[Normal[eventRecords[All, "Variable"]]]
  ];
```

The constructed rows are assembled into a dataset (with Dataset). The function ProcessComputationSpecification is used to convert a user-made specification dataset into a form used by ERTMon.

In[194]:= **wCompSpec = ProcessComputationSpecification[Dataset[compSpecRows]] [SortBy[#Variable &]]**

Out[194]=

	Variable	Explanation	MaxHistoryLength	AggregationIntervalLength	AggregationFunction	NormalizationScope	NormalizationFunction
Humidity.Mean	Humidity		5 184 000	172 800	Mean	Entity	Mean
Humidity.OutliersCount	Humidity		5 184 000	172 800	OutliersCount	Variable	None
Humidity.Range	Humidity		5 184 000	172 800	Range	Country	Mean
MaxTemperature.Mean	MaxTemperature		5 184 000	172 800	Mean	Entity	Mean
MaxTemperature.OutliersCount	MaxTemperature		5 184 000	172 800	OutliersCount	Variable	None
MaxTemperature.Range	MaxTemperature		5 184 000	172 800	Range	Country	Mean
Pressure.Mean	Pressure		5 184 000	172 800	Mean	Entity	Mean
Pressure.OutliersCount	Pressure		5 184 000	172 800	OutliersCount	Variable	None
Pressure.Range	Pressure		5 184 000	172 800	Range	Country	Mean
Temperature.Mean	Temperature		5 184 000	172 800	Mean	Entity	Mean
Temperature.OutliersCount	Temperature		5 184 000	172 800	OutliersCount	Variable	None
Temperature.Range	Temperature		5 184 000	172 800	Range	Country	Mean
WindSpeed.Mean	WindSpeed		5 184 000	172 800	Mean	Entity	Mean
WindSpeed.OutliersCount	WindSpeed		5 184 000	172 800	OutliersCount	Variable	None
WindSpeed.Range	WindSpeed		5 184 000	172 800	Range	Country	Mean

The computation specification is set to the monad with the function `ERTMonSetComputationSpecification`.

Alternatively, a computation specification can be created and filled-in as a CSV file and read into the monad. (Not described here.)

Grouping event records by entity-variable pairs

With the function `ERTMonGroupEntityVariableRecords` we group the event records by the found unique entity-variable pairs. Note that in the pipeline below we set the computation specification first.

```
In[195]:= p =  
  p⇒  
    ERTMonSetComputationSpecification[wCompSpec] ⇒  
    ERTMonGroupEntityVariableRecords;
```

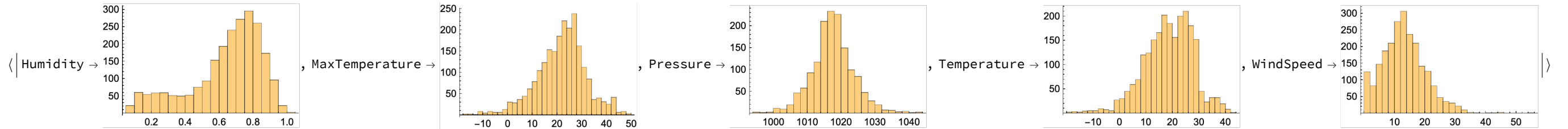
Descriptive statistics (per variable)

After the data is ingested into the monad and the event records are grouped per entity-variable pairs we can find certain descriptive statistics for the data. This is done with the general function `ERTMonComputeVariableStatistic` and the specialized function `ERTMonFindVariableOutlierBoundaries`.

```
In[196]:= p⇒ERTMonComputeVariableStatistic[RecordsSummary]⇒ERTMonEchoValue;  
  
» value: { { Humidity → { Min 0.07, 1st Qu 0.548, Mean 0.64175, Median 0.696, 3rd Qu 0.796, Max 1. },  
            { MaxTemperature → { Min -16.72, 1st Qu 16, Mean 21.6512, Median 22.78, 3rd Qu 27.78, Max 48 },  
            { Pressure → { Min 995.3, 1st Qu 1014.1, Median 1017.4, Mean 1017.52, 3rd Qu 1021, Max 1043.1 },  
            { Temperature → { Min -19.33, 1st Qu 12.78, Median 18.5323, Mean 19.11, 3rd Qu 25.17, Max 42.28 },  
            { WindSpeed → { Min 0, 1st Qu 8.33, Median 12.685, Mean 13.0895, 3rd Qu 17.22, Max 54.82 } } }
```

```
In[197]:= p ⇒ ERTMonComputeVariableStatistic ⇒ ERTMonEchoValue;
```

```
» value:
```



```
In[198]:= p ⇒ ERTMonComputeVariableStatistic[TakeLargest[#, 3] &] ⇒ ERTMonEchoValue;
```

```
» value: <| Humidity → {1., 1., 0.993}, MaxTemperature → {48, 48, 48}, Pressure → {1043.1, 1042.8, 1041.1}, Temperature → {42.28, 41.94, 41.89}, WindSpeed → {54.82, 44.63, 44.08} |>
```

Finding the variables outlier boundaries

The finding of outliers counts and fractions can be specified in the computation specification. Because of this there is a specialized function for outlier finding `ERTMonFindVariableOutlierBoundaries`. That function makes the association of the found variable outlier boundaries (i) to be the pipeline value and (ii) to be the value of context key “variableOutlierBoundaries”. The outlier boundaries are found using the functions of the package [Aap6].

If no argument is specified `ERTMonFindVariableOutlierBoundaries` uses the Hampel identifier (`HampelIdentifierParameters`).

```
In[199]:= p ⇒ ERTMonFindVariableOutlierBoundaries ⇒ ERTMonEchoValue;
```

```
» value: <| Humidity → {0.522536, 0.869464}, MaxTemperature → {14.2106, 31.3494}, Pressure → {1012.36, 1022.44}, Temperature → {9.88823, 28.3318}, WindSpeed → {5.96141, 19.4086} |>
```

```
In[200]:= Keys[p ⇒ ERTMonFindVariableOutlierBoundaries ⇒ ERTMonTakeContext]
```

```
Out[200]:= {eventRecords, entityAttributes, computationSpecification, entityVariableRecordGroups, variableOutlierBoundaries}
```

In the rest of document we use the outlier boundaries found with the more conservative identifier `SPLUSQuartileIdentifierParameters`.

```
In[201]:= p =
```

```
p ⇒
```

```
ERTMonFindVariableOutlierBoundaries[SPLUSQuartileIdentifierParameters] ⇒ ERTMonEchoValue;
```

```
» value: <| Humidity → {0.176, 1.168}, MaxTemperature → {-1.67, 45.45}, Pressure → {1003.75, 1031.35}, Temperature → {-5.805, 43.755}, WindSpeed → {-5.005, 30.555} |>
```

Conversion of event records to time series

The grouped event records are converted into time series with the function `ERTMonEntityVariableGroupsToTimeSeries`. The time series are aligned to the time point given as an argument. The argument can be: a date object, “MinTime”, “MaxTime”, or “None”. (“MaxTime” is the default.)

```
In[202]:= p ⇒
```




```
ERTMonEntityVariableGroupsToTimeSeries["MinTime"] ⇒
```

```
ERTMonEchoFunctionContext[#timeSeries[1, 3, 5]] &;
```

```
» <| {KMFL, Temperature} → TimeSeries[ Time: 0 to 31536 000  
Data points: 366], {KMFL, Pressure} → TimeSeries[ Time: 0 to 31536 000  
Data points: 347], {KMFL, WindSpeed} → TimeSeries[ Time: 0 to 31536 000  
Data points: 366] |>
```

Compare the last output with the output of the following command.

```
In[203]:= p =
p ==
ERTMonEntityVariableGroupsToTimeSeries["MaxTime"] ==
ERTMonEchoFunctionContext[#timeSeries[[{1, 3, 5}]] &];

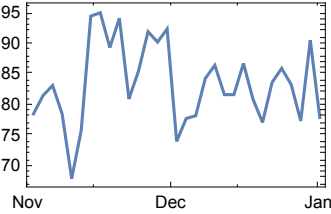
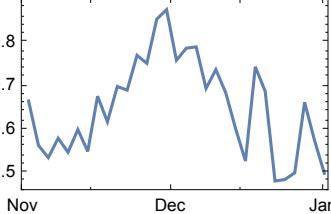
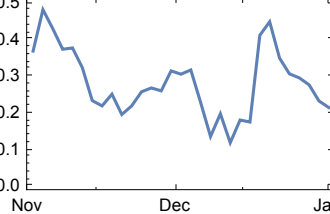
» <| {KMFL, Temperature} → TimeSeries[ Time: -31536000 to 0 Data points: 366], {KMFL, Pressure} → TimeSeries[ Time: -31536000 to 0 Data points: 347], {KMFL, WindSpeed} → TimeSeries[ Time: -31536000 to 0 Data points: 366] >|
```

Time series restriction and aggregation.

The main goal of ERTMon is to convert a diverse, general collection of event records into a collection of aligned time series over specified regular time grids.

The regular time grids are specified with the columns “MaxHistoryLength” and “AggregationIntervalLength” of the computation specification. The time series of the variables in the computation specification are restricted to the corresponding maximum history lengths and are aggregated using the corresponding aggregation lengths and functions.

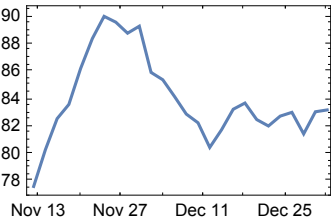
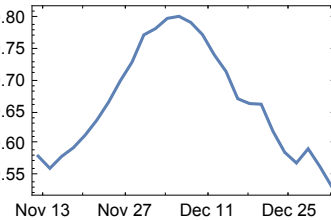
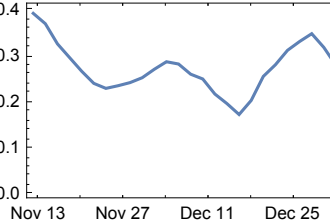
```
In[204]:= p =
p ==
ERTMonAggregateTimeSeries ==
ERTMonEchoFunctionContext[DateListPlot /@ #timeSeries[[{1, 3, 5}]] &];

» <| {KMFL, Humidity.Mean} → , {EGLC, Humidity.Mean} → , {KGEU, Humidity.Mean} →  >|
```

Application of time series functions

At this point we can apply time series modifying functions. An often used such function is moving average.

```
In[205]:= p ==
ERTMonApplyTimeSeriesFunction[MovingAverage[#, 6] &] ==
ERTMonEchoFunctionValue[DateListPlot /@ #[[{1, 3, 5}]] &];

» <| {KMFL, Humidity.Mean} → , {EGLC, Humidity.Mean} → , {KGEU, Humidity.Mean} →  >|
```

Note that the result is given as a pipeline value, the value of the context key “timeSeries” is not changed.

Normalization

With “normalization” we mean that the values of a given time series values are divided (normalized) with a descriptive statistic derived from a specified set of values. The specified set of values is given with the parameter “NormalizationScope” in computation specification.

At the normalization stage each time series is associated with an entity ID and variable.

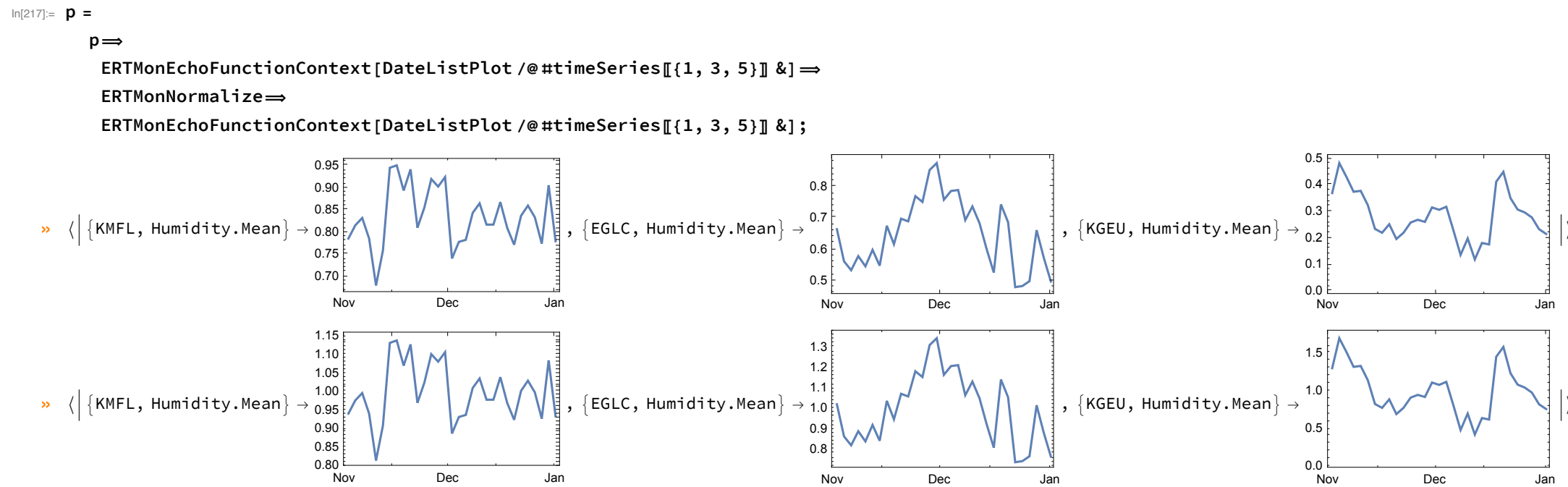
Normalization is done at three different scopes: “entity”, “attribute”, and “variable”.

Given a time series $T(i, var)$ corresponding to entity ID i and a variable var we define the normalization values for the different scopes in the following ways.

- Normalization with scope “entity” means that the descriptive statistic is derived from the values of $T(i, var)$ only.
- Normalization with scope attribute means that
 - from the entity attributes dataset we find attribute value that corresponds to i ,
 - next we find all entity ID's that are associated with the same attribute value,
 - next we find value of normalization descriptive statistic using the time series that correspond to the variable var and the entity ID's found in the previous step.
- Normalization with scope “variable” means that the descriptive statistic is derived from the values of all time series corresponding to var .

Note that the scope “entity” is the most granular, and the scope “variable” is the coarsest.

The following command demonstrates the normalization effect -- compare the y-axes scales of the time series corresponding to the same entity-variable pair.



Here are the normalization values that should be used when normalizing “unseen data.”

```
In[218]:= p=>ERTMonTakeNormalizationValues
```

```
Out[218]:= { {Humidity.Range, Country, USA} -> 0.0864597, {Humidity.Range, Country, UK} -> 0.066, {Humidity.Range, Country, Australia} -> 0.145968,
{MaxTemperature.Range, Country, USA} -> 2.85468, {MaxTemperature.Range, Country, UK} -> 78/31, {MaxTemperature.Range, Country, Australia} -> 3.28871,
{Pressure.Range, Country, USA} -> 2.08222, {Pressure.Range, Country, Australia} -> 3.33871, {Temperature.Range, Country, USA} -> 2.14411, {Temperature.Range, Country, UK} -> 1.25806,
{Temperature.Range, Country, Australia} -> 2.73032, {WindSpeed.Range, Country, USA} -> 4.13532, {WindSpeed.Range, Country, UK} -> 3.62097, {WindSpeed.Range, Country, Australia} -> 3.17226 }
```

Making contingency matrices

One of the main goals of ERTMon is to produce contingency matrices corresponding to the event records data.

The contingency matrices are created and stored as `SSparseMatrix` objects, [AAP7].

```
In[223]:= p =
p=>ERTMonMakeContingencyMatrices;
```

We can obtain an association of the contingency matrices for each variable-and-aggregation-function pair, or obtain the overall contingency matrix.

In[233]:=

p⇒ERTMonTakeContingencyMatrices

Dimensions /@%

Out[233]=

Humidity.Mean → SparseArray

Specified elements: 186

Dimensions: {6, 31}

Humidity.OutliersCount → SparseArray

Specified elements: 6

Dimensions: {6, 31}

Humidity.Range → SparseArray

Specified elements: 179

Dimensions: {6, 31}

MaxTemperature.Mean → SparseArray

Specified elements: 186

Dimensions: {6, 31}

MaxTemperature.OutliersCount → SparseArray

Specified elements: 0

Dimensions: {6, 31}

MaxTemperature.Range → SparseArray

Specified elements: 158

Dimensions: {6, 31}

Pressure.Mean → SparseArray

Specified elements: 121

Dimensions: {6, 31}

Pressure.OutliersCount → SparseArray

Specified elements: 3

Dimensions: {6, 31}

Pressure.Range → SparseArray

Specified elements: 105

Dimensions: {6, 31}

Temperature.Mean → SparseArray

Specified elements: 186

Dimensions: {6, 31}

Temperature.OutliersCount → SparseArray

Specified elements: 0

Dimensions: {6, 31}

Temperature.Range → SparseArray

Specified elements: 180

Dimensions: {6, 31}

WindSpeed.Mean → SparseArray

Specified elements: 173

Dimensions: {6, 31}

WindSpeed.OutliersCount → SparseArray

Specified elements: 1

Dimensions: {6, 31}

WindSpeed.Range → SparseArray

Specified elements: 165

Dimensions: {6, 31}

Out[234]=

Humidity.Mean → {6, 31}, Humidity.OutliersCount → {6, 31}, Humidity.Range → {6, 31}, MaxTemperature.Mean → {6, 31}, MaxTemperature.OutliersCount → {6, 31},

MaxTemperature.Range → {6, 31}, Pressure.Mean → {6, 31}, Pressure.OutliersCount → {6, 31}, Pressure.Range → {6, 31}, Temperature.Mean → {6, 31},

Temperature.OutliersCount → {6, 31}, Temperature.Range → {6, 31}, WindSpeed.Mean → {6, 31}, WindSpeed.OutliersCount → {6, 31}, WindSpeed.Range → {6, 31} |>

In[237]:=

p⇒ERTMonTakeContingencyMatrix;

MatrixPlot[%, ImageSize → 700]

Out[238]=

In[231]:=

RowNames[smat]

Out[231]=

{EGLC, KGEU, KMDW, KMFL, KNIP, WM095866}

Larger example pipeline

The pipeline shown in this section utilizes all workflow functions of ERTMon. The used weather data and computation specification are described above.

ERTMonUnit[] ⇒

ERTMonSetEventRecords[eventRecords] ⇒

ERTMonSetEntityAttributes[entityAttributes] ⇒

ERTMonEchoDataSummary ⇒

ERTMonSetComputationSpecification[compSpec] ⇒

ERTMonGroupEntityVariableRecords ⇒

ERTMonFindVariableDistributions[Histogram[#1, ImageSize → Small] &] ⇒

Out[] = ERTMonEchoFunctionValue["Variable distributions:", #1 &] ⇒

ERTMonFindVariableOutlierBoundaries[SPLUSQuartileIdentifierParameters] ⇒

ERTMonEchoFunctionValue["Outlier boundaries:", #1 &] ⇒

ERTMonEntityVariableGroupsToTimeSeries["MaxTime"] ⇒

ERTMonAggregateTimeSeries ⇒

ERTMonMakeContingencyMatrices ⇒

ERTMonEchoFunctionValue["Contingency matrices:", (MatrixPlot[ToSSparseMatrix[#1], ImageSize → Small] &) /@ #1 &]

initialize the monad pipeline

ingest event records

ingest event attributes

show data summary

ingest computation specification

group records per {EntityID, Variable}

find variable distributions

echo found distributions

find outliers boundaries for each variable

echo found outliers boundaries

align the event records and create time series

aggregate the time series

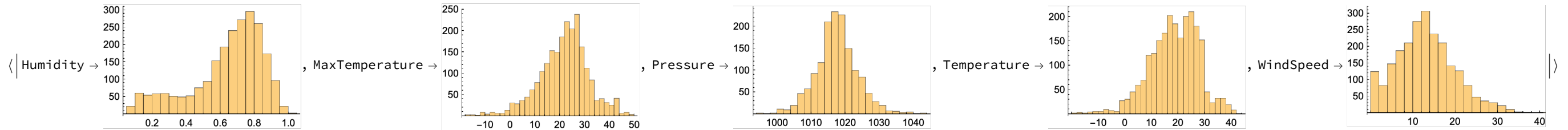
(according to the computation specification)

make contingency matrices

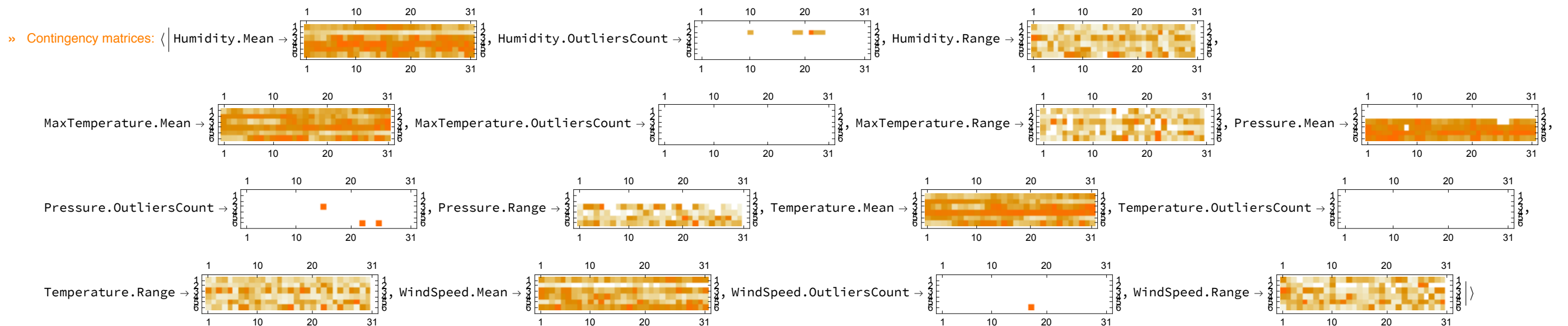
plot contingency matrices

» Data summary: $\langle \text{eventRecords} \rightarrow \left\{ \begin{array}{l} \text{KNIP} \quad 1967 \quad \text{Jacksonville} \quad 1967 \quad \text{Min} \quad 3\,689\,539\,200 \\ \text{WM095866} \quad 1965 \quad \text{Melbourne} \quad 1965 \quad \text{1st Qu} \quad 3\,705\,955\,200 \\ \text{KMDW} \quad 1802 \quad \text{Chicago} \quad 1802 \quad \text{Mean} \quad 3.71474 \times 10^9 \\ \text{KMFL} \quad 1798 \quad \text{Miami} \quad 1798 \quad \text{Median} \quad 3\,715\,113\,600 \\ \text{KGEU} \quad 1572 \quad \text{Peoria} \quad 1572 \quad \text{3rd Qu} \quad 3\,723\,235\,200 \\ \text{EGLC} \quad 1456 \quad \text{London} \quad 1456 \quad \text{Max} \quad 3\,739\,392\,000 \end{array} \right\}, \text{entityAttributes} \rightarrow \left\{ \begin{array}{l} \text{EGLC} \quad 2 \quad \text{USA} \quad 4 \\ \text{KGEU} \quad 2 \quad \text{Australia} \quad 1 \\ \text{KMDW} \quad 2 \quad \text{Chicago} \quad 1 \\ \text{KMFL} \quad 2 \quad \text{Jacksonville} \quad 1 \\ \text{KNIP} \quad 2 \quad \text{London} \quad 1 \\ \text{WM095866} \quad 2 \quad \text{Melbourne} \quad 1 \\ \text{(Other)} \quad 3 \end{array} \right\} \rangle$

» Variable distributions:



» Outlier boundaries: $\langle \text{Humidity} \rightarrow \{0.176, 1.168\}, \text{MaxTemperature} \rightarrow \{-1.67, 45.45\}, \text{Pressure} \rightarrow \{1003.75, 1031.35\}, \text{Temperature} \rightarrow \{-5.805, 43.755\}, \text{WindSpeed} \rightarrow \{-5.005, 30.555\} \rangle$



References

Packages

- [AAp1] Anton Antonov, State monad code generator Mathematica package, (2017), MathematicaForPrediction at GitHub.
URL: <https://github.com/antononcube/MathematicaForPrediction/blob/master/MonadicProgramming/StateMonadCodeGenerator.m> .
- [AAp2] Anton Antonov, Monadic tracing Mathematica package, (2017), MathematicaForPrediction at GitHub.
URL: <https://github.com/antononcube/MathematicaForPrediction/blob/master/MonadicProgramming/MonadicTracing.m> .
- [AAp3] Anton Antonov, Monadic Event Records Transformations Mathematica package, (2018), MathematicaForPrediction at GitHub.
URL: <https://github.com/antononcube/MathematicaForPrediction/blob/master/MonadicProgramming/MonadicEventRecordsTransformations.m> .
- [AAp4] Anton Antonov, MathematicaForPrediction utilities, (2014), MathematicaForPrediction at GitHub.
URL: <https://github.com/antononcube/MathematicaForPrediction/blob/master/MathematicaForPredictionUtilities.m> .
- [AAp5] Anton Antonov, Cross tabulation implementation in Mathematica, (2017), MathematicaForPrediction at GitHub.
URL: <https://github.com/antononcube/MathematicaForPrediction/blob/master/CrossTabulate.m> .
- [AAp6] Anton Antonov, Implementation of one dimensional outlier identifying algorithms in Mathematica, (2013), MathematicaForPrediction at GitHub.
URL: <https://github.com/antononcube/MathematicaForPrediction/blob/master/OutlierIdentifiers.m> .
- [AAp7] Anton Antonov, SSparseMatrix Mathematica package, (2018), MathematicaForPrediction at GitHub.
URL: <https://github.com/antononcube/MathematicaForPrediction/blob/master/SSparseMatrix.m> .
- [AAp8] Anton Antonov, Monadic contextual classification Mathematica package, (2017), MathematicaForPrediction at GitHub.
URL: <https://github.com/antononcube/MathematicaForPrediction/blob/master/MonadicProgramming/MonadicContextualClassification.m> .
- [AAp9] Anton Antonov, Weather event records data Mathematica package, (2018), MathematicaForPrediction at GitHub.
URL: <https://github.com/antononcube/MathematicaForPrediction/blob/master/Misc/WeatherEventRecords.m> .

Documents

- [AA1] Anton Antonov, Monad code generation and extension, (2017), MathematicaForPrediction at WordPress.
URL: <https://mathematicaforprediction.wordpress.com/2017/06/23/monad-code-generation-and-extension> .
- [AA1a] Anton Antonov, Monad code generation and extension, (2017), MathematicaForPrediction at GitHub.
- [AA2] Anton Antonov, A monad for classification workflows, (2018), MathematicaForPrediction at WordPress.
URL: <https://mathematicaforprediction.wordpress.com/2018/05/15/a-monad-for-classification-workflows> .