Многие файлы были некорректно отформатированы: элемент, содержащий замер по климатическому фактору и элемент, содержащий пропуск в измерениях (равный 9.99, 99.9 или — в некоторых случаях — 1.0), часто не разделяются пробелом. При обработке данных человеком такая ошибка легко отслеживается, но при обработке на компьютере такие два элемента рассматриваются как один, что приводит к ошибкам. Поясним на примере:

рассмотрим данные по фактору "нижняя облачность" за 1973 год:

1.2 2.2 1.4 1.3 6.4 7.8 7.0 7.8 4.599.9 .6 .8 1973

В примере данные за октябрь и ноябрь не разделены пробелом и при загрузке данных в MATLAB они будут рассматриваться как один элемент.

Отформатировать текст можно с помощью различных инструментов: консольных текстовых редакторов (например, sed), текстовых процессоров, входящих в состав офисных пакетов (MS Word, OO Writer и пр.) и т.д. Из-за простоты редактирования и специфики задачи коррекция таких ошибок производится с помощью скрипта, написанного на языке программирования Perl.

Группировка

Исходная структура данных не подходит для биоклиматических расчетов, ведущихся по отдельным станциям: данные должны группироваться так же по станциям, но каждый файл, соответствующий станции, должен содержать все климатические параметры, сгруппированные по месяцам: строки соответствуют годам, а столбцы отсортированы по 8 климатическим параметрам; порядок расположения параметров строго определен (в соответствии со структурой входных данных модели теплопотерь). При этом первая группа из 8 столбцов соответствует январю, вторая – февралю и далее по месяцам года. Таким образом, исходные двумерные массивы данных должны быть преобразованы в трехмерные.

Для решения задачи группировки была разработана программа на встроенном языке MATLAB.

Работу программы можно разбить на 3 этапа.

1. Производится загрузка данных в MATLAB. Данные из текстового формата преобразуются в числовой и располагаются в массиве ячеек. Каждая ячейка, в свою очередь, содержит M ячеек, в которых хранятся двумерные массивы данных по замерам очередного климатического фактора: строки соответствуют годам, столбцы – месяцам.

2. После загрузки данных следует избавиться от столбцов, содержащих избыточную информацию – нумерацию по годам. Затем данные сортируются таким образом, чтобы первые M столбцов получившихся массивов (соответствующих станциям) содержали данные по всем замерам всех факторов за январь, следующие M – за февраль и так далее по месяцам года.

3. Отсортированные данные записываются на диск.

Пропуски в данных

Перед тем, как будут окончательно сформированы массивы входных данных для модели энергопотерь, необходимо заполнить имеющиеся пропуски в метеоданных: в связи с тем, что данные собирались в течение длительного промежутка времени и на большой территории, для многих станций отсутствуют значения некоторых климатических факторов для отдельных месяцев или лет наблюдений. Пропуски в публикациях данных наблюдений месячного усреднения могут встречаться по нескольким причинам. К ним относятся: временное закрытие станции, неполнота наблюдений за месяц, браковка наблюдений. В исходных данных они помечаются как 99.9 или 999.9.

Одним из важных для биометрических вычислений климатических факторов является плотность снега. К сожалению, данные по этому параметру зачастую отсутствуют – начиная с 1990 года на большинстве станций замеры по нему отсутствуют вообще. Для того, чтобы установить зависимость между плотностью снега и другими имеющимися в наличии метеоданными, необходимо построить модель, определить, какие из факторов оказывают влияние на плотность снега и получить параметры модели.

Простейший вариант – линейный вариант модели:

$$x = a_0 + a_1 T + a_2 W + a_3 H + a_4 D + a_5 Cl + a_6 Cc + a_7 P \ .$$

где x – плотность снега, T – температура, W – скорость ветра, H – влажность, D – глубина снежного покрова, Cl – нижняя облачность, Cc – общая облачность, P – количество осадков, $a_0..a_7$ – настраиваемые параметры модели.

Для оценки настраиваемых параметров плотность снега должна рассчитываться на различных наборах метеоданных. Каждый набор - значения климатических факторов для некоторой метеостанции, конкретного месяца и года. Для каждого набора определяется невязка, равная разности между рассчитанной и фактической плотностью снега. В качестве критерия при настройке параметров выбирается среднеквадратичное отклонение по всем наборам метеоданных.

Счет можно организовать разным способом, но наиболее простой и эффективный – матричный вариант. При этом сразу можно рассчитать невязки, включив в правую часть уравнения фактическую плотность снега со знаком минус.

В этом случае критерий настройки параметров K будет определяться так:

$$D = AF\,, \quad S = D^T D\,, \quad K = \sqrt{\frac{S}{n}}\,,$$

где A – вектор-столбец параметров, F – матрица метеоданных, D – вектор-столбец невязок, S – сумма их квадратов, K - критерий настройки параметров – скаляр, n – число данных в наборе (соответствует годам, для которых имеются замеры).

Матрица параметров включает дополнительный единичный столбец для учета фактической плотности снега. Матрица метеоданных будет включать кроме указанных выше факторов также столбец плотности снега и единичный столбец для свободного члена a_0.

Подбор параметров модели производится с помощью встроенного в пакет AnyLogic оптимизатора OptQuest. Оптимизатор OptQuest для решения задачи оптимизации комбинирует эвристические методы, нейронные сети и математическую оптимизацию, которая состоит из нескольких последовательных прогонов модели с различными значениями оптимизационных параметров и нахождении оптимальных параметров для нашей задачи.

Для использования оптимизатора была построена модель в AnyLogic, соответствующая приведенной выше математической модели.

Из-за неоднородности данных перед их использованием в модели необходимо их нормировать.

Нормировка данных

Так как исходные метеоданные, участвующие в расчетах, неоднородны и их величины могут отличаться на порядок, оценить по коэффициенту степень влияния отдельного фактора на модель не представляется возможным. Поэтому для дальнейшего

использования в моделях данные необходимо пронормировать. Нормировка будет проводится следующим образом:

$$X_{norm} = \frac{x_i - x_{min}}{|x_{max} - x_{min}|}.$$

В результате набор данных будет представлять из себя числа от 0 до 1:

$0 < M(i) < 1$.

Модель плотности снега

В ходе расчетов с помощью OptQuest выяснилось, что наилучшее значение целевого функционала достигается на наборах данных T, H и D (температура воздуха, влажность воздуха и глубина снежного покрова соответственно), т.е. линейный вариант модели принял вид:

$$x = a_0 + a_1 T + a_2 H + a_3 D \cdot$$

В таком случае матрица F будет выглядеть так:

$$F = \begin{matrix} 1\ T_1\ H_1\ D_1\ X_{f1} \\ \dots \\ 1\ T_N\ H_N\ D_N\ X_{fN}. \end{matrix}$$

В среде AnyLogic матрица F формируется с помощью метода set из единичного вектора и значений переменных – векторов T, H и D.

Вектор B:

$$B = \begin{pmatrix} a_0 \\ a_T \\ a_H \\ a_D \\ -1 \end{pmatrix}.$$

Вектор-столбец невязок *Delta* определяется как произведение матрицы F и переменной – вектора B.

Для контроля корректности полученных параметров данные за месяц, по которому производится настройка, разбиваются на 2 группы: одна будет использоваться для их настройки, вторая – для контроля. Например, для настройки берутся данные по январю за годы 1970–1986, для контроля – 1986–1990 за этот же месяц.

Для определения оптимальных значений параметров модели в среде AnyLogic создается оптимизационный эксперимент:

– в качестве целевого функционала указывается СКО рассчитанного в ходе эксперимента значения плотности снега и фактических значений, полученных от метеостанций; при расчете СКО используется метод transpose класса matrix для получения значений квадратов невязок;

– в качестве параметров выступают элементы вектора B, минимальное и максимальное значение устанавливается в пределах [–1;1], в ходе повторных запусков оптимизационного эксперимента эти значения корректируются.

Для повышения точности были введены нелинейные члены.

В этом случае модель приняла вид:

$$x = a_0 + a_1 T + a_2 H + a_3 D + a_4 T^2 + a_5 H^2 + a_6 D^2 + a_7 TH + a_8 TD + a_9 HD \ .$$

Вектор B:

$$B = \begin{pmatrix} a_0 \\ a_T \\ a_H \\ a_D \\ a_{T^2} \\ a_{H^2} \\ a_{D^2} \\ a_{TH} \\ a_{TD} \\ a_{HD} \\ -1 \end{pmatrix} .$$

Это усложняет задачу оптимизации, поэтому необходимо значительно увеличить количество прогонов модели. Но в результате достигается бóльшая точность модели – значение целевого функционала меньше на порядок.

При анализе результатов становится понятно, что для расчета плотности снега для декабря и января можно рекомендовать модель, настроенную по данным января, для февраля, марта, апреля и мая – модель, настроенную по данным февраля (для оценки адекватности применения той или иной модели вычислялось СКО среднеарифметического плотности снега и ее расчетных значений для различных наборов месяцев). Параметры модели, полученные при ее настройке на данных за указанные месяцы, приведены в таблице.

Январь		Февраль	
a_0	0,2486	a_0	0,2679
a_T	0,091	a_T	0,026
a_H	0,063	a_H	0,0968
a_D	-0,076	a_D	0,0842
a_{T^2}	0,0509	a_{T^2}	0,0009
a_{H^2}	-0,0128	a_{H^2}	-0,1056
a_{D^2}	0,2574	a_{D^2}	-0,0428
a_{TH}	-0,037	a_{TH}	0,0117
a_{TD}	-0.2161	a_{TD}	-0,0104
a_{HD}	0,0391	a_{HD}	0,0003

Выводы

Компьютерные эксперименты показали, что максимальная абсолютная ошибка в расчетах ($\left|X_p - X_\phi\right|$, где X_p – плотность снега, полученная в результате расчета, X_ϕ – плотность снега, полученная со станций) равна 0,06 (минимальная – на порядок мень-

ше), что составляет около 20% от среднего значения плотности снега для контрольного месяца. Это немало, но для практического применения важна не абсолютная ошибка, а какая-либо статистическая характеристика, например среднеквадратическое отклонение, а оно составляет около 7,5% от среднего значения рассчитываемого параметра.

Из-за того, что наборов данных достаточно мало (всего 8), желательно осуществлять настройку параметров модели и расчет пропущенных замеров по наборам данных близкорасположенных станций. Также необходимо учитывать сезонные особенности: например весной могут быть оттепели, которые невозможно учесть по среднемесячной температуре, но которые могут влиять на плотность снега.

Литература

1. **Мордовин В. Ю., Михайлов В. В.** Модель энергозатрат животных и климат // Труды СПИИРАН / РАН. С.-Петербург. ин-т информатики и автоматизации; Под общ. ред. Р. М. Юсупова. Т. 2. Вып. 2. СПб.: Наука, 2005. С. 407–417.
2. **Дрейпер Н., Смит Г.** Прикладной регрессионный анализ: В 2-х кн. Кн. 1 /Пер. с англ. 2-е изд., перераб и доп. М.: Финансы и статистика, 1986.
3. **Тулупьев А. Л., Николенко С. И., Сироткин А. В.** Байесовские сети: логико-вероятностный подход. СПб.: Наука, 2006. 607 с.
4. **Тулупьев А. Л., Сироткин А. В., Николенко С. И.** Синтез согласованных оценок истинности утверждений в интеллектуальных информационных системах // Изв. высш. учебн. заведений: Приборостроение. 2006. №7. 20–26 с.