

УДК 004.67

Методы усреднения ошибки и многомерной линейной регрессии для статистической обработки численных прогнозов погоды

*Дикопалова Т.С., студент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»*

*Научный руководитель: Шляева А.В., к.т.н, доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
irudakov@bmstu.ru*

Существующие численные модели прогнозирования погоды принимают в качестве входных данных состояние атмосферы, полученное в результате обработки данных метеонаблюдений, собранных по всему миру. Неидеальность входных данных прогнозирующей модели и её программной реализации могут приводить к систематической ошибке в прогнозе [1].

Задача обработки выходных данных модели прогнозирования, или постпроцессинга, состоит в выявлении статистических зависимостей между прогнозируемыми значениями переменных модели и измеренными значениями этих переменных. Найденные зависимости могут быть использованы как для коррекции выходных переменных модели прогнозирования, так и для предсказания значений переменных, явно не прогнозируемых моделью [2].

Помимо непосредственно корректируемых прогнозов и данных о наблюдениях данные методы принимают на вход и другие параметры, от которых будет зависеть, насколько получится улучшить прогноз. Например, необходимо знать минимальный объем данных, на которых достаточно обучить метод, чтобы далее применять его в реальных условиях. В литературе обычно даются общие рекомендации к значению данной величины [4]. Кроме того, при описании методов постпроцессинга зачастую не оговаривают, следует ли при корректировании прогноза в данной точке учитывать ошибки прогнозов в соседних точках [4, 5, 6]. Представляет интерес получение более или менее точных значений периода обучения и радиуса окрестности точки, которую имеет смысл учитывать, поскольку они являются необходимыми для применения методов постпроцессинга в практических задачах.

Применение методов постпроцессинга для корректирования прогнозов

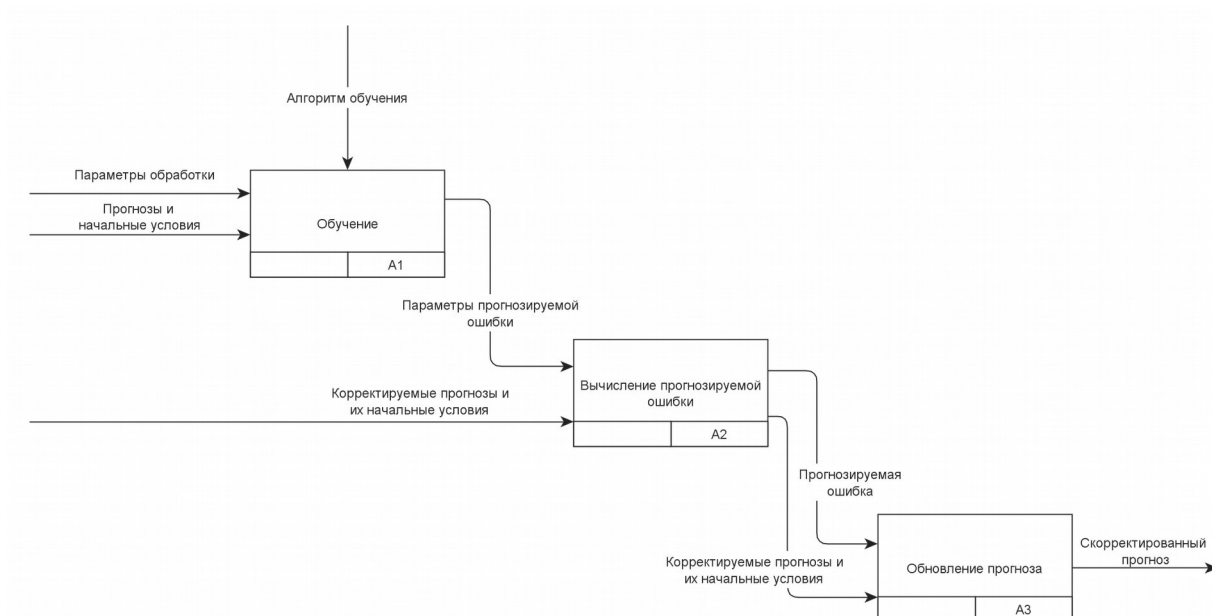


Рис. 1. Схема работы метода постпроцессинга

Общий подход методов постпроцессинга к обработке результатов модели прогнозирования и корректированию прогноза представлен на рисунке 1. По форме обучения можно выделить два типа алгоритмов постпроцессинга: адаптивные и неадаптивные. В неадаптивных методах обучение происходит один раз, в адаптивных проводится повторно по мере поступления новых данных. Главным недостатком неадаптивных методов является то, что они требуют «замораживания» моделей прогнозирования [4]. Среди недостатков адаптивных методов можно выделить то, что за счет относительно небольшого периода обучения, их точность может быть не велика, а также они более восприимчивы к резким изменениям данных, которые могут быть вызваны, например, ошибкой измерений.

Прогнозы некоторой величины в разных географических точках на момент времени t обычно рассматривают в виде двухмерной сетки, в качестве шага которой берут градусы долготы и широты.

Наиболее простой метод частичного устранения систематической ошибки модели — усреднение ошибки за определённый период и вычитание её из текущего прогноза [5]. Тогда ошибку прогноза в i -ой точке сетки можно вычислить по формуле 1:

$$\varepsilon_{i\text{cp}}(t_1, t_2) = \sum_{t=t_1}^{t_2} \frac{\varepsilon_i}{(t_2 - t_1) + 1}, i = 1..n \quad (1)$$

где t_1, t_2 — начало и конец периода обучения, n — общее количество точек сетки.

При применении многомерной линейной регрессии для постпроцессинга прогнозов в качестве факторов могут выступать значения начальных условий или ошибок прогноза переменных модели прогнозирования, астрономические или географические параметры на время прогноза, а также их предыдущие измерения. При вычислениях оперируют вектором коэффициентов линейной регрессии и матрицей начальных значений факторов, которая содержит как начальные значения факторов, так и начальное значение корректируемой переменной модели прогнозирования [4].

Приведённые выше методы могут применяться как адаптивно, так и неадаптивно. Для адаптивного применения необходимо после получения начальных условий по каждому новому прогнозу повторно производить обучение по сдвинутому вперёд окну данных. При этом размер окна данных не должен быть меньше периода прогноза.

Учёт окрестности точки в методах постпроцессинга

Вычисление ошибки прогноза в данной точке с учётом ошибок прогнозов в соседних точках можно выполнить по формуле 2.

$$\varepsilon_i = \frac{\sum_{j \in S_i} \varepsilon_j w_{ij}}{\sum_{j \in S_i} w_{ij}}, \quad (2)$$

где S_i — окрестность i -ой точки. В качестве функции весов рассматривается функция расстояния между двумя точками. Приведённую формулу удобно применять для учёта окрестности точки в методе усреднения ошибки.

Учесть окрестность i -ой точки в методе многомерной линейной регрессии можно несколькими способами. Например, при заполнении начальных значений факторов можно для каждого фактора занести прогнозы как в данной точке, так и в точках её окрестности. При таком подходе не учитывается степень влияния значений в точке из окрестности на значение величины в данной точке. Это можно учесть, если при обучении вычислять вектора коэффициентов регрессии в каждой точке отдельно, а при обработке прогноза вычислять ошибку в каждой точке по формуле 2.

Структура программного обеспечения для исследования методов статистической обработки численных прогнозов

Для обучения методов постпроцессинга и последующего тестирования их в данной работе использовались данные о прогнозах, предоставляемые в открытом доступе организацией Earth System Research Laboratory (ESRL)¹. Выходные данные модели прогнозирования представляются в виде архивов формата NetCDF (.nc).

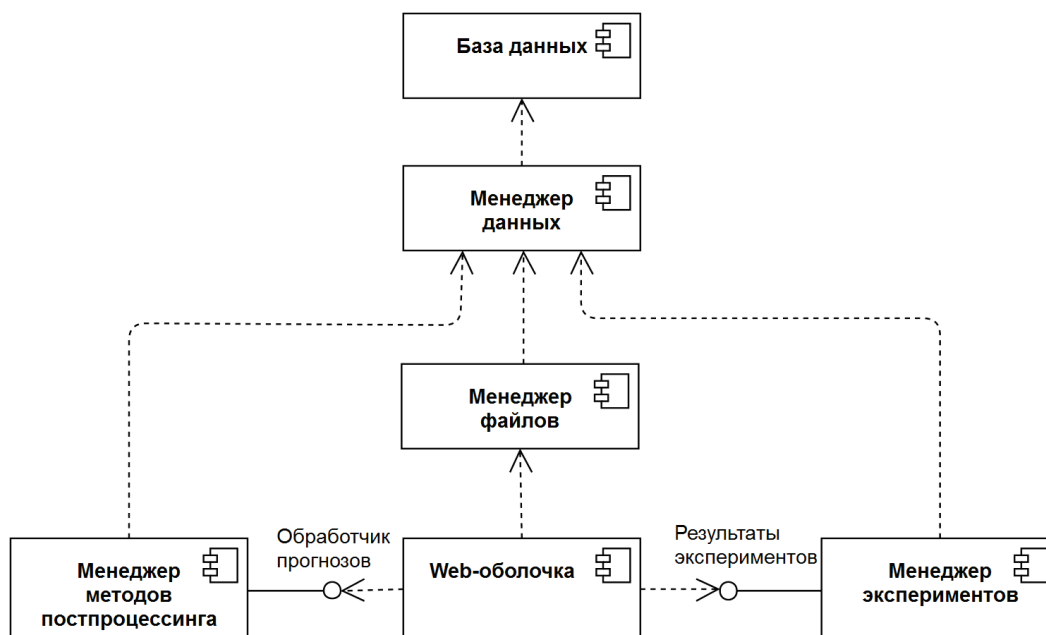


Рис. 2. Основные компоненты программного обеспечения

Для обеспечения быстрого доступа к данным, было реализовано внутреннее хранилище данных в виде базы данных, предусматривающее хранение некоторой мета-информации. Для предоставления возможности исследования других реализаций методов постпроцессинга в системе предусмотрена динамическая загрузка сторонних реализаций. Структура программного обеспечения представлена на рисунке 2.

Поскольку данные по прогнозам погоды достаточно объёмны, выполнение расчётов может занимать длительное время. По этой причине ПО реализовано как удалённый сервис, доступ к которому предоставляется через Web-оболочку.

В условиях больших объёмов данных как по прогнозам, так и по начальным условиям, необходимо обрабатывать данные пошагово, разделяя их на независимые части и обрабатывая последовательно или параллельно. Для метода усреднения ошибки этого можно добиться дроблением данных по времени. В случае метода линейной регрессии для подсчёта коэффициентов в каждой точке необходимы все имеющиеся на период обучения измерения каждой из величин-факторов, вследствие чего необходимо разделять сетку на некоторые участки, выполнять для них вычисления отдельно.

¹ <http://esrl.noaa.gov/psd/forecasts/reforecast2>

Определение окрестности точки

Для определения влияния ошибок в окрестности точки на ошибку в данной были проведены опыты с прогнозами температуры на высоте 2 метров. Для метода усреднения ошибки опыты проводились с периодом обучения семь месяцев и диапазоном корректируемых данных — следующие три месяца. По результатам, приведённым в таблице 1, видно, что радиус учитываемой окрестности точки не влияет на результаты применения метода.

Таблица 1

Опыты с окрестностью точки для метода усреднения ошибки

№	Радиус окрестности	Средняя ошибка до обработки	Средняя ошибка после обработки
1.	4.0	0.1605	0.0109 (Уменьшилась на 93%)
2.	0.0		0.0109 (Уменьшилась на 93%)

Для метода многомерной линейной регрессии период обработки был равен одному месяцу, а в качестве фактора использовалась влажность на высоте 2 метров. По результатам этих опытов, представленных в таблице 2, можно сделать вывод о том, что при выборе фактора влажность для коррекции прогноза температуры на высоте 2 метров учёт окрестности точки ухудшает результаты обработки методом многомерной линейной регрессии.

Опыты с окрестностью точки для метода многомерной линейной регрессии

№	Период обучения	Окрестность точки	Средняя ошибка до обработки	Средняя ошибка после обработки
1.	02.01.2012 — 16.01.2012	4.0	0.195	0.975 (Увеличилась в 5 раз)
2.	02.01.2012 — 16.01.2012	0.0		0.311 (Увеличилась на 59%)

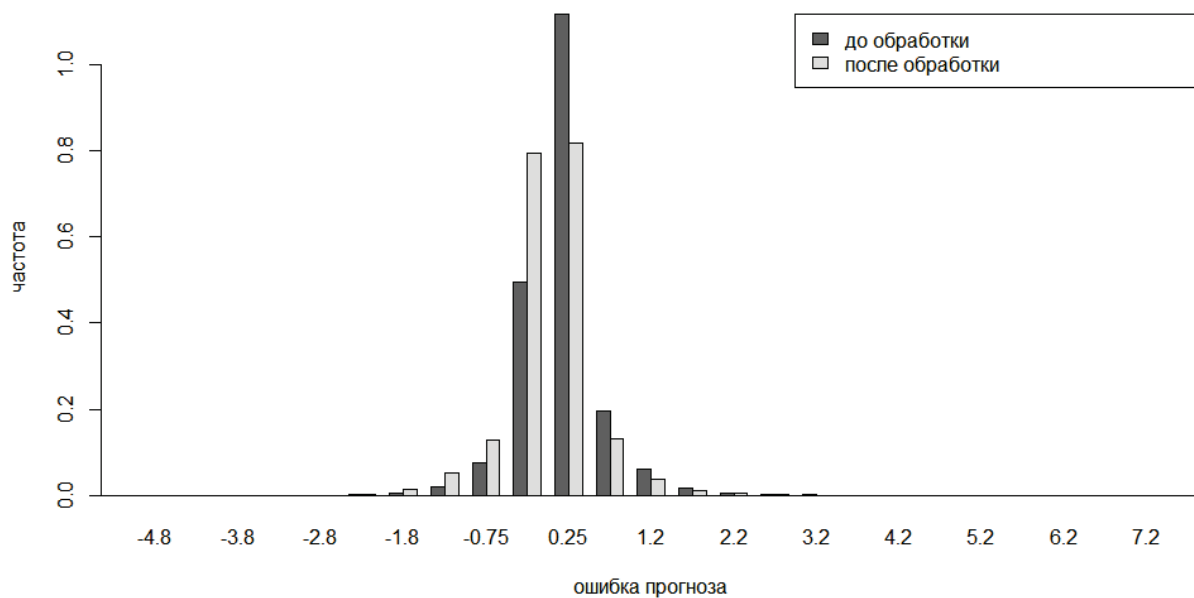
Определение длительности обучения

Для определения длительности обучения проводились опыты без учёта окрестности точки, поскольку, как было показано выше, учёт ошибок в соседних точках не улучшает результаты применения методов. Методы применялись неадаптивно. Корректируемым прогнозом был выбран прогноз температуры на высоте 2 метров. Для метода многомерной линейной регрессии был взят один фактор — влажность на высоте 2 метров. После обучения методы применялись для обработки прогнозов на период 02.01.2013 — 01.06.2013.

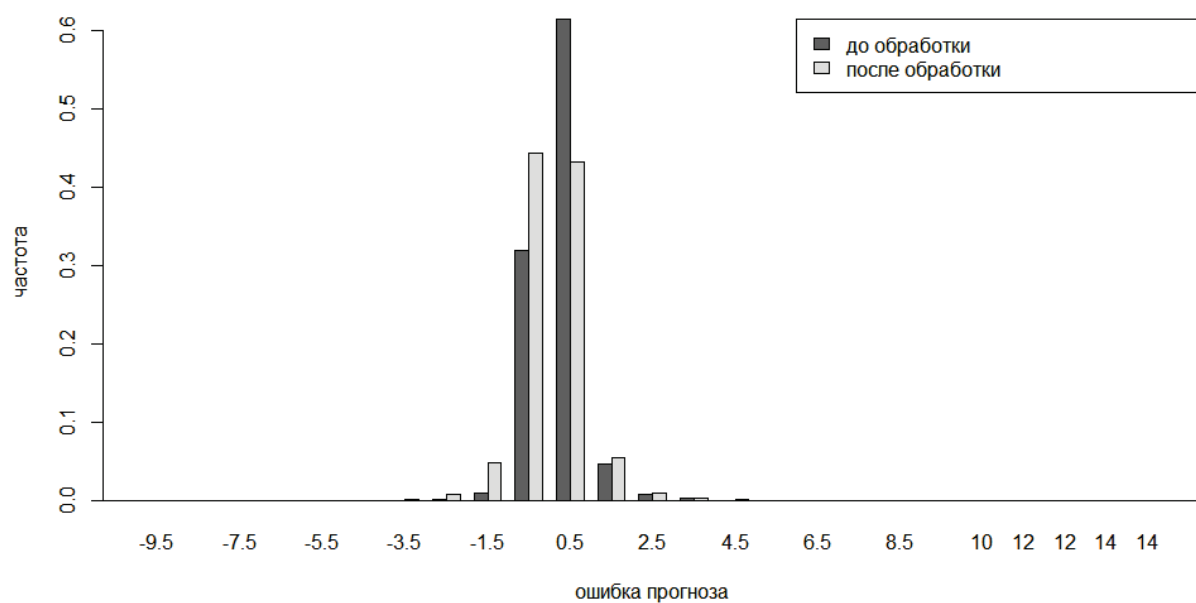
Результаты опытов приведены в таблице 3. На рисунках 3, а и 3, б приведены гистограммы распределения ошибки до и после обработки прогноза для опытов 3 и 6. На основании этих данных можно сделать следующие выводы.

Метод усреднения ошибки показывает хороший результат при небольшой длительности периода обучения. При возрастании длительности результаты его применения ухудшаются.

Метод многомерной линейной регрессии показывает хороший результат обработки при длительности периода обучения от приблизительно семи месяцев и до года. На периоде обучения в один год результаты применения начинают незначительно ухудшаться.



а)



б)

Рис. 3. а) Распределение ошибки до и после применения метода усреднения ошибки с периодом обучения 1 месяц; б) Распределение ошибки до и после применения метода многомерной линейной регрессии с периодом обучения 7 месяцев.

Опыты с окрестностью точки для метода усреднения ошибки

№	Метод	Период обучения	Средняя ошибка до обработки	Средняя ошибка после обработки
1.	Усреднение ошибки	02.01.2012 — 01.01.2013 (1 год)	0.148	0.005
2.	Усреднение ошибки	02.01.2012 — 31.07.2012 (6 месяцев)	0.161	0.011
3.	Усреднение ошибки	02.01.2012 — 01.02.2012 (1 месяц)	0.148	-0.003
4.	Усреднение ошибки	02.01.2012 — 15.02.2012 (2 недели)	0.148	-0.016
5.	Многомерная линейная регрессия	02.01.2012 — 01.01.2013 (1 год)	0.148	0.031
6.	Многомерная линейная регрессия	02.01.2012 — 31.07.2012 (7 месяцев)	0.161	0.020
7.	Многомерная линейная регрессия	12.01.2012 — 01.02.2012 (1 месяц)	0.148	0.325

Определение множества факторов линейной регрессии

Для определения множества факторов, влияющих на прогноз температуры на высоте 2 метра с периодом прогноза 24 часа, проверялись различные комбинации факторов. Опыты проводились с периодом обучения семь месяцев, данные обрабатывались на периоде в четыре месяца. Окрестность точки при этом не учитывалась.

Сначала было определено, какие факторы оказывают влияние на ошибку в прогнозе температуры, а какие нет. По результатам, представленным в таблице 4, видно, что на температуру наибольшее влияние оказывают влажность на высоте 2 метра и ожидаемые осадки, плотность облачного покрова и потенциальная завихрённость оказывают чуть меньшее влияние, а объёмная влажность почвы не влияет вовсе.

С величинами, которые оказывают влияние, были проведены дополнительные опыты. Судя по их результатам, метод многомерной линейной регрессии лучше всего корректирует значение температуры при множестве факторов, состоящем из влажности на высоте 2 метра и ожидаемых осадков.

Эксперименты с множеством факторов

№	Множество факторов	Средняя ошибка до обработки	Средняя ошибка после обработки
1.	Влажность на высоте 2 метра	0,1273	0.0201
2.	Ожидаемые осадки	0.1273	0.0219
3.	Объёмная влажность почвы	0.1273	-2.3928
4.	Плотность облачного покрова	0.1273	0.1248
5.	Потенциальная завихрённость	0.1273	0.1261
6.	Плотность облачного покрова Влажность на высоте 2 метра Ожидаемые осадки	0,1273	0.0050
7.	Влажность на высоте 2 метра Ожидаемые осадки	0,1273	-0.0008
8.	Плотность облачного покрова Ожидаемые осадки	0.1273	0.0155

По результатам экспериментов, проведённых с помощью разработанного ПО, для метода усреднения ошибки и линейной регрессии можно сделать следующие выводы.

Для метода усреднения ошибки не имеет разницы, применяется ли он с учётом окрестности точки или без учёта окрестности. Следовательно, для уменьшения количества вычислений можно применять его, всегда учитывая ошибку по всей сетке. Для метода многомерной линейной регрессии учёт окрестности ошибки точки ухудшает результат.

Метод усреднения ошибки показывает наилучшие результаты при периоде обучения около 1 месяца, из чего следует, что он хорошо подходит для адаптивного обучения, так как оптимальный период обучения небольшой.

Метод многомерной линейной регрессии по одному фактору показывает наилучшие результаты обработки при периоде обучения около 6 месяцев, следовательно, в чистом виде не должен применяться адаптивно.

На ошибку прогноза температуры на высоте 2 метра больше всего влияют ошибки в прогнозе влажности на высоте 2 метра и в прогнозе ожидаемых осадков.

Список литературы

1. Phillips N.A. The general circulation of the atmosphere: a numerical experiment // Quarterly Journal of the Royal Meteorological Society. 1956. Vol. 82. P. 123–154. DOI: 10.1002/qj.49708235202.
2. Glahn, H.R., Dale A.L. The Use of Model Output Statistics (MOS) in Objective Weather Forecasting // J. Appl. Meteor. 1972 . Vol. 11. P. 1203–1211. DOI: 10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2
3. Guidelines on Ensemble Prediction Systems and Forecasting. World Meteorological Organization. Available at:
https://googledrive.com/host/0BwdvoC9AeWjUazhkNTdXRXUzOEU/wmo_1091_en.pdf,
accessed 15.05.2014.
4. Kalnay E. Atmospheric Modeling, Data Assimilation and Predictability. Cambridge: Cambridge University Press, 2003. 341 p.
5. McCollor D., Stull R.B. Hydrometeorological Accuracy Enhancement via Postprocessing of Numerical Weather Forecasts in Complex Terrain // Wea. Forecasting. 2008. Vol. 23. P. 131–144. DOI: 10.1175/2007WAF2006107.1
6. Wilks D.S., Hamill T.M. Comparison of Ensemble-MOS Methods Using GFS Reforecasts // Mon. Wea. Rev. 2007. Vol. 135. P. 2379–2390. DOI: 10.1175/MWR3402.1