Использование теста Гренджера при прогнозировании временных рядов*

А. П. Мотренко
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

Работа посвящена исследованию возможностей применения теста Гренджера в прогнозировании временных рядов. В основе теста Гренджера лежат статистические тесты и использование линейных регрессионных моделей. Исследуется зависимость качества прогноза от порядка модели, способа обработки данных. В вычислительном эксперименте приводятся результаты работы алгоритма на различных временных рядах: стационарных, нестационарных, с обратной связью, независимых по Гренджеру.
Ключевые слова: *тест Гренджера, casual connectivity, выбор порядка регрессионной модели.*

Введение

Назовем временным рядом последовательность значений некоторой величины $\mathbf{x}(t)$, измеренной через равные промежутки времени. Основываясь на этих данных, а также исследуя другие временные ряды, можно спрогнозировать значения ряда $\mathbf{x}(t)$ в будущем. В частности, возникают ситуации, когда история других рядов лучше помогает сделать прогноз, чем история самого $\mathbf{x}(t)$. Определить, существует ли зависимость такого рода, дает возможность тест Гренджера – статистический метод, предложенный Клайвом Гренджером (Clive Granger) в 60-х годах [1].

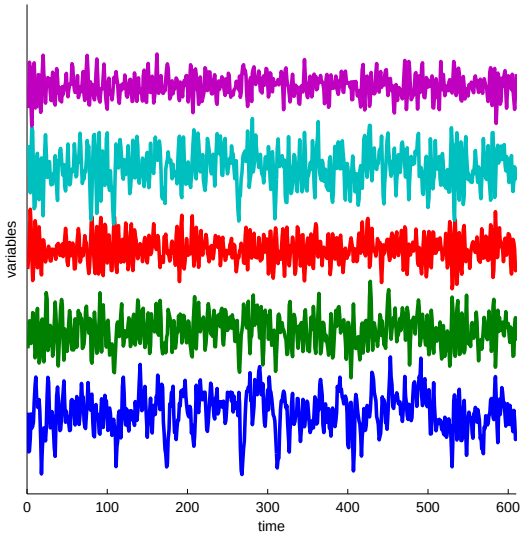

Рис. 1. Вид исследуемых рядов.

Постановка задачи

Использование теста Гренджера подразумевает прогнозирование с помощью линейных регрессионных моделей. Пусть $\mathbf{x}_1(t)$ и $\mathbf{x}_2(t)$ — исследуемые временные ряды, тогда пример

Научный руководитель В. В. Стрижов

такой модели

$$\mathbf{x}_1(t) = \sum_{j=1}^{p} a_{11}(j)\mathbf{x}_1(t-j) + \sum_{j=1}^{p} a_{12}(j)\mathbf{x}_2(t-j) + E_1(t),$$

$$\mathbf{x}_2(t) = \sum_{j=1}^{p} a_{21}(j)\mathbf{x}_1(t-j) + \sum_{j=1}^{p} a_{22}(j)\mathbf{x}_2(t-j) + E_2(t).$$

Здесь p – количество предыдущих значений, принимаемых во внимание, матрица $A(j)$ с коэффицентами $a_{ik}(j)$ содержит веса узлов, а $E_1(t)$ и $E_1(t)$ – ошибки прогнозирования. Будем считать, что $\mathbf{x}_1$ следует из $\mathbf{x}_2$, если ошибка прогнозирования $\mathbf{x}_1(t)$ уменьшается при включении в модель значений ряда $\mathbf{x}_2$ (то есть если коэфты $A_{12,j}$ заметно отличаются от нуля). Результат чувсвителен к изменению параметра p; эта зависимость исследуется в работе.

Тест Гренджера применим к рядам, обладающим постоянными (не зависящими от времени) матожиданием и дисперсией. Если исследуемый ряд не обладает этими свойствами, необходимо привести его к соответсвующему виду. Предполагается провести исследование зависимости результата от способа преобразования данных.

Для оценки качества прогнозирования выделим контрольную выборку длины m и построим функционал

$$Q = \sum_{i=1}^{m} (\tilde{\mathbf{x}}(i) - \mathbf{x}(i))^2,$$

где $\tilde{\mathbf{x}(i)}$ — спрогнозированное значение элемента выборки, $\mathbf{x}(i)$ — его истинное значение.

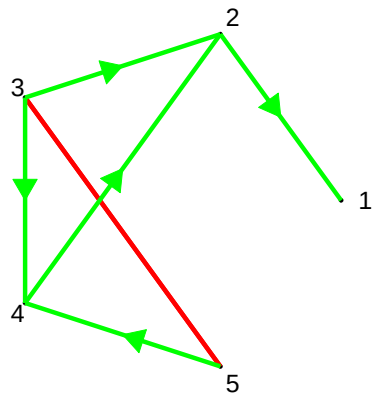

Рис. 2. Схема зависимостей между рядами. Цифры соответствуют номерам рядов, например, ряд $\mathbf{x}_2$ зависит от $\mathbf{x}_3$ и $\mathbf{x}_4$. Красной линией обозначена двусторонняя связь.

Пути решения

Подготовка данных. Прежде всего, необходимо удостовериться, что ряды стационарны. Для оценки этого свойства можно использовать ADF тест [2], основанный на нулевой гипотезе об отсутствии стационарности или, дополнительно к ADF, KPSS тест [3],

предполагающий ее наличие. Если проверка дала отрицательный результат, необходимо модифицировать ряды. Существуют различные способы:

— дифференцирование ряда — т.е. переход непосредственно от значений к их изменениям. Эта операция увеличивает вероятность успеха и может повторяться несколько раз, однако каждая итерация затрудняет интерпретацию полученных данных, поэтому в данной работе дифференцирование проводится лишь один раз. С данными, для которых однократное дифференцирование не приводит к положительному результату (то есть не удается получить стационарный ряд), алгоритм не работает.

— метод окна – использовать лишь часть известного ряда. Подход основан на идее, что чем короче ряд тем больше он похож на стационарный.

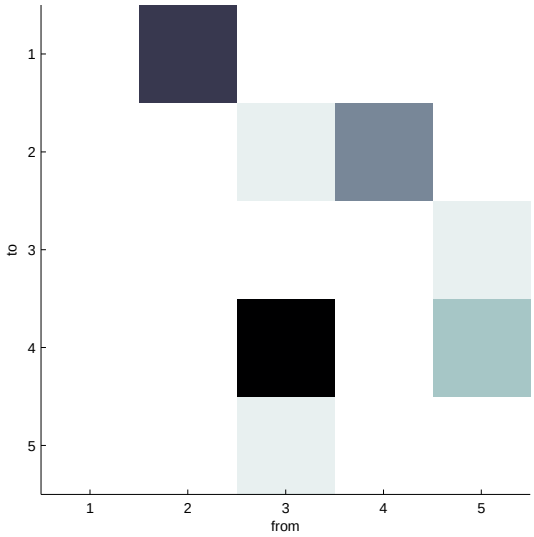

Рис. 3. Альтернативный способ представления результатов. Здесь закрашенная клетка на пересечении строки и столбца означает зависимость элемента слобца от элемента строки. Например, 2 влечет за собой 1.

Выбор параметров. Ключевым параметром при использовании теста Гренджера является порядок модели (порядок лагирования), т.е. количество предыдущих измерений (значений ряда), учтенных при прогнозировании очередного значения. Если этот параметр не может быть выбран на основе априорного знания, то могут быть использованы информационные критерии Акаике (Akaike information criterion, AIC) [4] или Байеса (Bayesian information criterion, BIC) [5], позволяющие сравнивать модели с различным числом параметров. Рассмотрим их подробнее.

В данной работе рассматривается только линейная регрессия, поэтому критерий Акаике может быть представлен в виде:

$$AIC = 2p + m \ln \frac{Q}{m}.$$

Аналогичный вид в случае линейной регрессии имеет и байесовский критерий, однако функция штрафа за сложность модели (т.е за ее порядок) здесь жестче:

$$BIC = p \ln m + m \ln \frac{Q}{m}.$$

В обоих случаях наилучшей модели соответсвует минимальное значение критерия. Важно, что модели сравниваюстя по выборкам одинаковой длины.

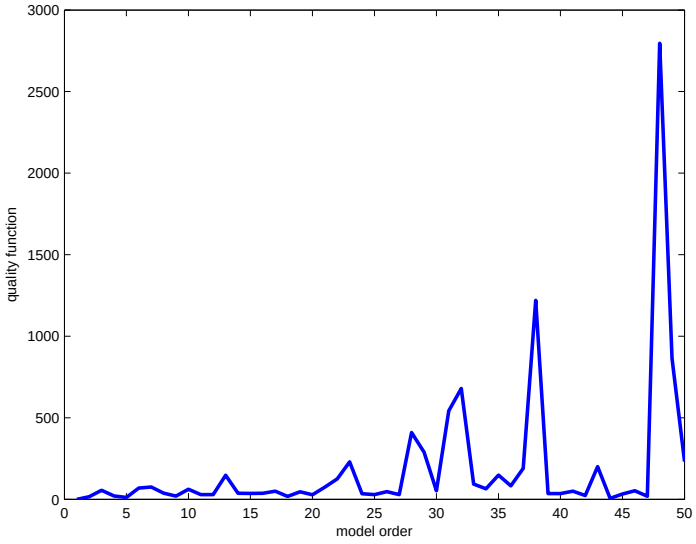

Рис. 4. Зависимость относительного функционали качества от порядка модели p для $\mathbf{x}_2$.При p больших 5 значение ошибок значительно возрастает.

Прогнозирование временных рядов. В данном эксперименте регрессия строится следующим образом: пусть $\mathbf{x}_{K\times T}$ — матрица, строки которой содержат элементы временных рядов, а столбцы соответсвуют моментам времени. В ней выделим последние p столбцов. Чтобы определить матрицу коэффицентов регрессии, составим матрицу R размером $(T-p)\times Kp$, полученную из $\mathbf{x}$, со строками r_i вида:

$$r_i = \begin{pmatrix} \mathbf{x}_1(T-2p+k-1) \\ \mathbf{x}_1(T-2p+k) \\ \dots \\ \mathbf{x}_1(T-p+k-1) \\ \mathbf{x}_2(T-2p+k-1) \\ \dots \\ \dots \\ \mathbf{x}_K(T-p+k-1) \end{pmatrix}^T . \tag{1}$$

Тогда

$$\tilde{\mathbf{x}_i} = R\beta_i,$$

где $\tilde{\mathbf{x}_i}$ — вектор, составленный из элементов ряда $\mathbf{x}_i$, от $T-p$ до T, b_i – - вектор коэффицентов регрессии. Таким образом можем определить β_i:

$$\beta_i = \tilde{\mathbf{x}_i} R^{-1}.$$

Пусть мы хотим выяснить, зависит ли ряд $\mathbf{x}(t)$ от $y(t)$. Действуем по следующей схеме [6, гл.17]:

1. Прогнозируем ряд $\mathbf{x}$ линейной регрессией с порядком p. При этом используем значения этого ряда и любых других известных рядов, кроме $y(t)$. По полученным данным вычисляем
$$RSS_R = \sum_{i=1}^{n} (\tilde{\mathbf{x}}(t-i) - \mathbf{x}(t-i))^2.$$
Здесь RSS_R — обозначение для residual sum of squares (restricted), $\tilde{\mathbf{x}}(t-i)$ — по-прежнему, спрогнозированное значение элемента ряда, $\mathbf{x}(t-i)$ — его истинное значение.
2. Прогнозируем ряд $\mathbf{x}(t)$ линейной регрессией, но уже с использованием элементов ряда $y(t)$. По полученным данным вычисляем
$$RSS_{UR} = \sum_{i=1}^{n} (\hat{\mathbf{x}}(t-i) - \mathbf{x}(t-i))^2.$$
Здесь RSS_{UR} — обозначение для residual sum of squares (unrestricted).
3. Нулевая гипотеза: элементы Y не участвуют в регрессии.
4. Определим величину F следующим образом:
$$F = \frac{(RSS_R - RSS_{UR})/p}{RSS_{UR}/(n-m)}.$$
Обозначения: p — количество элементов ряда Y, задействованных в регрессии п.2, совпадающее с порядком модели; m — длина контрольной выборки.
5. Если вычисленное значение F превосходит некоторое критическое значение, отвергаем нулевую гипотезу, т.е $\mathbf{x}$ зависит от y.
6. Аналогичные действия проводим на случай существования обратной связи.

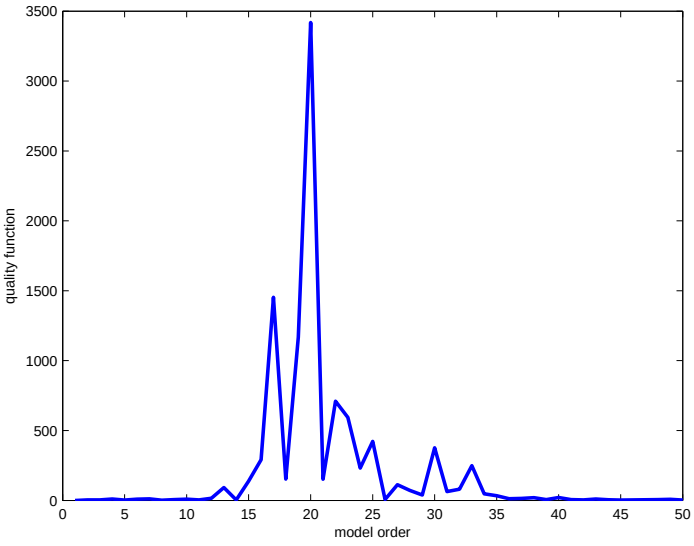

Рис. 5. Аналогичная зависимость для $\mathbf{x}_3$.Здесь качество прогнозирования ухудшается уже при p больших 2, что и соответсвует значениям, найденным с помощью BIC и AIC.

Вычислительный эксперимент

Работа алгоритма на модельных данных. В ходе эксперимента использовались синтетические данные(рис. 1), состоящие из пяти рядов, для получения которых поступим следующим образом: сгенерируем матрицу $\mathbf{x}_{5\times T}$, каждый элемент которой — нормальная случайная величина; затем переопределим каждый столбец, начиная с пятого, чтобы каждый элемент такого столбца удовлетворял формуле

$$\begin{gathered}\mathbf{x}_1(i) = 1.6\mathbf{x}_1(i-1) + 0.65\mathbf{x}_2(i-2),\\ \mathbf{x}_2(i) = 1.5\mathbf{x}_2(i-1) - 0.3\mathbf{x}_2(i-2) - 0.3\mathbf{x}_3(i-4) + 0.6\mathbf{x}_4(i-1),\\ \mathbf{x}_3(i) = 1.8\mathbf{x}_3(i-1) - 0.7\mathbf{x}_3(i-2) - 0.1\mathbf{x}_5(i-3),\\ \mathbf{x}_4(i) = 1.5\mathbf{x}_4(i-1) + 0.9\mathbf{x}_3(i-2) + 0.4\mathbf{x}_5(i-2),\\ \mathbf{x}_5(i) = 1.7\mathbf{x}_5(i-1) - 0.5\mathbf{x}_5(i-2) - 0.2\mathbf{x}_3(i-1).\end{gathered}$$

Таким образом каждый ряд определяется не только своей историей, но и прошлыми значениями других рядов. Например, ряд $\mathbf{x}_5$ зависит от $\mathbf{x}_1$, а изменения рядов $\mathbf{x}_3$ и $\mathbf{x}_4$ повлекут за собой изменение $\mathbf{x}_2$. Эта зависимость продемонстрирована на рисунках 2 и 3.

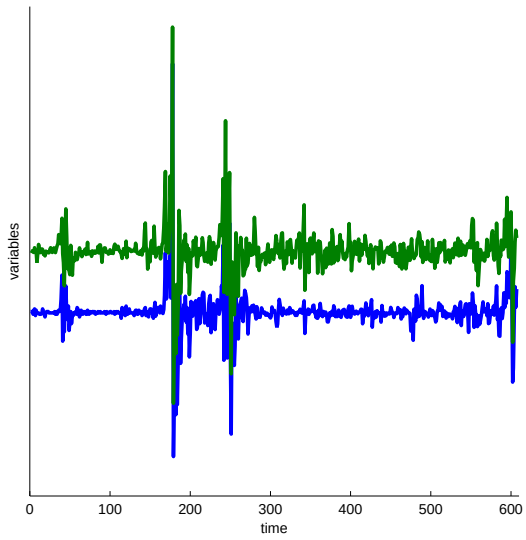

Рис. 6. Вид зависимостей цен на сахар от времени: синим в — Европе, зеленым — в США.

В данном случае все ряды оказались стационарными, дополнительная обработка не потребовалась. Критерии AIC и BIC дали оптимальное значение порядка модели, равное 2. Этот результат можно проверить, постороив зависимость функционала качества от порядка модели. Для наглядности будем использовать так называемый относительный функционал качества:

$$\tilde{Q} = \sum_{i=1}^{n} \left(\frac{\tilde{\mathbf{x}}(i) - \mathbf{x}(i)}{\mathbf{x}(i)} \right)^2 .$$

Рис. 4 и 5 демонстрируют, что при увеличении порядка модели ($p > 2$) функционал резко возрастает, в соответствии с критериями Аикаке и Байеса.

Работа алгоритма на реальных данных. В работе так же использовались реальные данные — цены на различные виды товаров за каждый месяц с 1960 по 2010 год [9]. В частности, были исследованы

1. зависимость цен на сахар в США от цен на сахар в Европе 6.
2. связь между ценами на природный газ в США и Европе и ценами на энергию (World Bank).

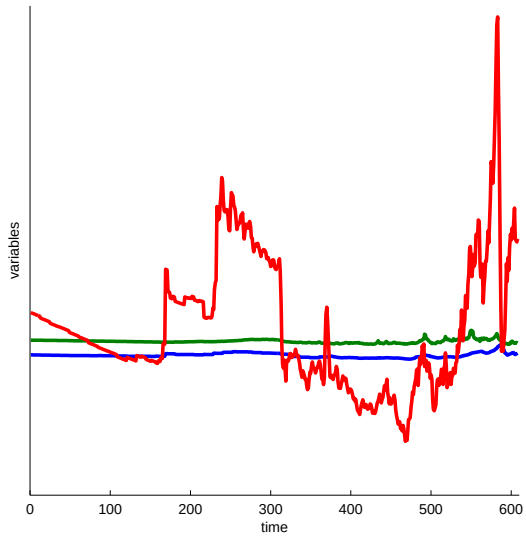

Рис. 7. Вид зависимостей для данных (2). Синим и зеленым цвет соответствует ценам на газ в Европе и Америке, красным — ценам на энергию.

В первом случае алгоритм выявил зависимость вида "цены на сахар в Европе влияют на цены на сахар в США". Такого результата можно было ожидать, изучив вид рис. 6. Чтобы убедиться в правильности результата, построим также зависимость относительных функционалов качества от порядка модели прогноза цен в США с использованием данных о ценах (назовем его unrestricted) в Европе и без (restricted) 8. Видно, что оценка, полученная при учете дополнительной информации, более точна. масштабе, видно что использование вспомогательных данных привело к более точному прогнозу.

В заключение данного раздела отметим, что из среди приведенных примером стационарностью не обладали временные ряды во втором случае. Из указанных в разделе "Подготовка данных-методов обработки наиболее действенным оказалось дифференцирование. В действительности, ни в одном из рассмотренных случаев метод окна не привел к положительному результату. Что касается применения алгоритма к заведомо нестационарным данным, возможны как точный прогноз, так и полное несоответствие с действительностью.

Заключение

В работе рассмотрена возможность использования теста Гренджера при составлении прогнозов. Исследована зависимость качества работы алгоритма от различных параметров, выбор которых проанализирован с точки зрения качества прогнозирования; также рассмотрены способы обработки входных данных. Анализ проводился на модельных и реальных данных.

Необходимый для построения вычислительного эксперимента код можно найти найти на сайте: `https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/GrangerForecasting/`

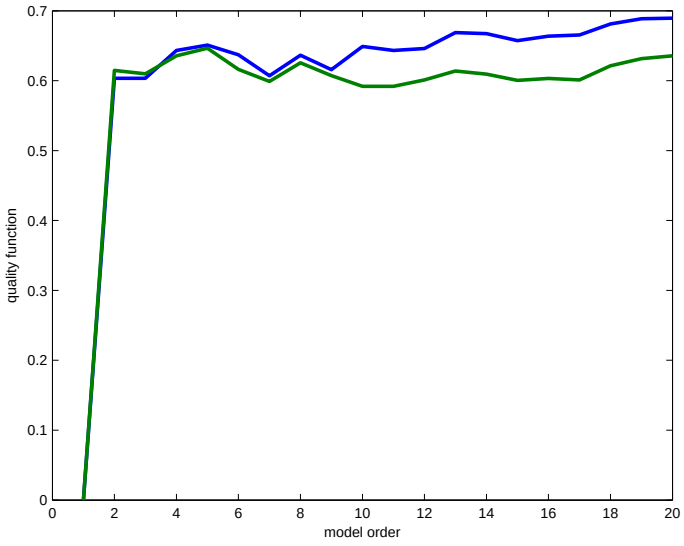

Рис. 8. Зависимость относительных функционалов качества от порядка модели: синим цветом — для прогнозирования по собственной истории(цены на сахар в США), зеленым — по истории цен на сахар в США и Европе.

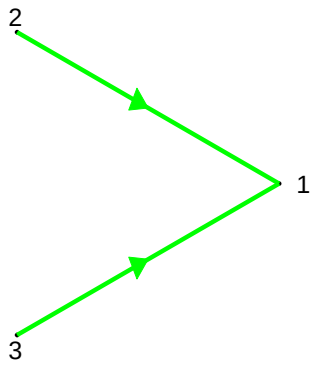

Рис. 9. 1 — цены на природный газ в Европе,2 — в США, 3 — цены на энергию. Здесь, например, 1 зависит от 3 и 2.

Результаты, полученные во втором случае, представлены на рисунке 9. Чтобы продемонстрировать справедливость результата, на этот раз представим на одном рисунке графики зависимостей исходного ряда и его предсказанных значений от времени 10. Как и в предыдущем случае, построим для ряда 1 unrestricted и restricted прогнозы. Из рис. 11, отражающего ту же зависимость, но в другом

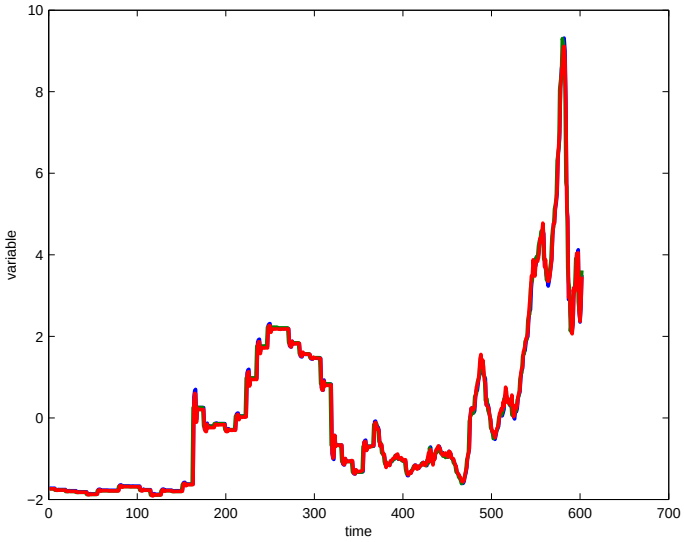

Рис. 10. Прогноз для цен на естественный газ в Европе; зеленым цветом обозначены исходные данные, синим — полученные в результате работы алгоритма.Ввиду нестационарности данных, на графиках отображены не сами значения исследуемых величин, а их изменения.

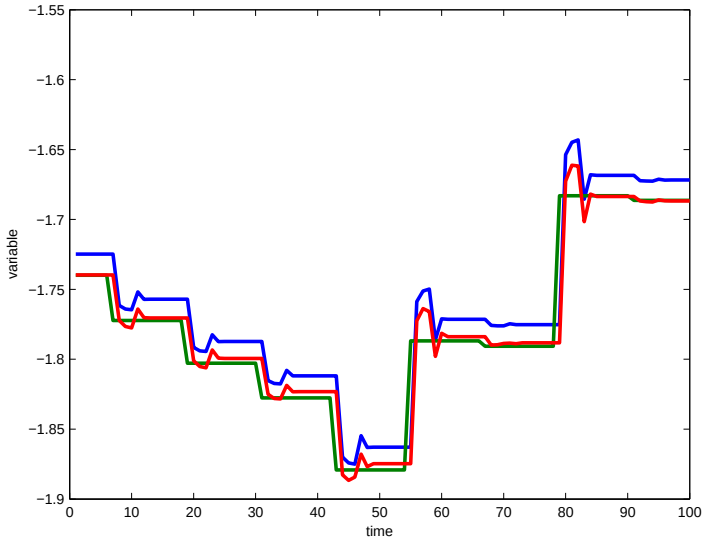

Рис. 11. Прогноз цен на естественный газ в Европе, в увеличенном масштабе. Здесь зеленым цветом обозначены исходные данные, синим — полученные в результате прогнозирования с использованием вспомогательных данных, красным — прогноз по истории самого ряда.

Литература

[1] C. W. J. Granger. *Investigating Causal Relations by Econometric Models and Cross-spectral Methods*, Econometrica, vol.37,424 - 432, 1969.

[2] J. D. Hamilton. *Time series analysis*, Princeton University Press, 1994.

[3] D. Kwiatskowski & Peter C. B. Phillips and Peter Schmidt & Yongcheol Shin. *Testing the null hypothesis of stationarity against the alternative of a unit root*, Journal of Econometrics, vol. 54, 159-178, 1992.

[4] H. Akaike. *A new look at the statistical model identification*, IEEE Trans. Autom.Control, vol. 19, 716-723, 1974.

[5] *Информационный критерий Байеса на MacnineLearning.ru*

[6] D. Gujarati. *Basic Econometrics, 4th ed*, The McGraw-Hill Companies, 2004.

[7] A. K. Seth. *A MATLAB toolbox for Granger causal connectivity analysis*,Journal of Neuroscience Methods,vol. 186, 262 - 273, 2010.

[8] A. K. Seth. *Granger causality*, Scolarpedia, 2(7), 2007.

[9] `https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/TSForecasting/TimeSeries/Sources/tsEarthquakesArkansas.csv`