

Исследование загрязненности воздуха в двух районах Лондона с использованием интеллектуального метода линейной регрессии в системе R

Беляев Алексей Андреевич

*Приамурский государственный университет им. Шолом-Алейхема
студент*

Баженов Руслан Иванович

*Приамурский государственный университет им. Шолом-Алейхема
к.п.н., доцент, зав. кафедрой информационных систем, математики и
правовой информатики*

Аннотация

В данной статье исследуется зависимость загрязнённости воздуха в двух регионах Лондона с использованием интеллектуального метода линейной регрессии в системе R. Зависимость использованных данных является хорошей.

Ключевые слова: R, система, RStudio, регрессия, линейная, интеллектуальный метод, Лондон, воздух, загрязнённость, регион, воздух

The study of air pollution in two districts of London using the intelligent linear regression method in the R system

Belyaev Alexey Andreevich

*Sholom-Aleichem Priamursky State University
student*

Bazhenov Ruslan Ivanovich

*Sholom-Aleichem Priamursky State University
Candidate of pedagogical sciences, associate professor, Head of the Department
of Information Systems, Mathematics and Legal Informatics*

Abstract

This article examines the dependence of air pollution in two regions of London using the intelligent linear regression method in the R system. The dependence of the used data is good.

Keywords: R, system, RStudio, regression, linear, intelligent method, London, air, pollution, region, air.

Большинство фирм, корпораций, учреждений, предприятий имеют огромное количество данных, которые каждый год меняются в лучшую или худшую сторону и чтобы спрогнозировать эти данные и узнать какими они

будут в последующих годах, то используются интеллектуальные методы статистического анализа данных. В зависимости, какие данные собираются анализировать, то нам необходимо выбрать для этого эффективный интеллектуальный метод, который даст наиболее существенный и правдивый прогноз самих данных на будущий период и покажет в них зависимость, что лучше, а что хуже будет.

В статье В. В. Стрижкова и Р. А. Сологуба применили алгоритм выбора нелинейных регрессионных моделей [1]. И. Ю. Глухих разработал модель экспресс-анализа финансовой состоятельности организаций на базе методов многомерного регрессионного анализа [2]. В статье Л.М.Бугаевского и Г.Г.Прохорова рассматривается разработка методики составления карт взаимосвязи с использованием корреляционного и регрессионного анализов [3]. В.С.Баджанов и Е.А.Матушевская применили корреляционно-регрессионный анализ для анализа себестоимости продукции на примере ГУП АО "Севастопольский Винодельческий завод" [4]. В статье А.А.Манцаева проведен анализ долговременных тенденций производительности труда в РК: корреляционно-регрессионный анализ [5]. О. Ю. Легкодух, Д. Д. Капустина и Т. А. Кокодей провели анализ и прогноз динамики курса доллара, используя инструментарий регрессионного анализа [6]. В статье Н. Х. Рашитова был проведен анализ эффективности структуры экономики на основе корреляционно-регрессионного анализа [7]. R. Boukezzoula, S. Galichet и D. Coquin в своей статье рассматривали анализ параметрических интервальных регрессионных методологий в соответствии с онтологическими и эпистемическими видениями интервалов [8]. В статье F. O. de França исследуется математическое выражение, описывающее взаимосвязь набора объясняющих переменных с измеряемой переменной [9]. I. Georgiev, D. I. Harvey, Stephen J. Leybourne, A.M. и R. Taylor рассмотрели тесты на структурные изменения, основанные на статистических данных типа SupF и Cramer-von-Mises Andrews и Nyblom [10].

Целью данной статьи является исследование зависимости загрязненности воздуха в двух районах Лондона с использованием интеллектуального метода линейной регрессии в системе R.

Большинство мира всегда интересуется вопросом загрязненности воздуха, в каком-то регионе или области. Для этого существует очень много разных данных, которые собирают ученые, ну или просто заинтересованные люди. Данные каждый год или месяц меняются в разных направлениях. В зависимости, какие данные собираются анализировать, то необходимо выбрать для этого эффективный интеллектуальный метод, который даст наиболее существенный и правдивый прогноз самих данных на будущий период и покажет в них зависимость, что лучше, а что хуже будет.

Данные, которыми мы пользовались в данном исследовании являются данные о загрязнении воздуха в пригороде и в центре города Лондон. Данные были взяты с начало января 2008 года по 18 апреля 2008 года. В данных указаны все имеющиеся примеси веществ, которые более или менее

находились в воздухе в тот момент времени. Примеси, которые были выявлены в момент исследования:

1. Диоксид азота (NO₂).
2. Оксид азота (NO).
3. Озон (O₃).
4. PM₁₀.
5. PM_{2.5}.
6. Диоксид серы.

Все примеси измеряются в (микрограммах на метр кубический) единица объёма или концентрации.

Для анализа и выявления зависимостей данных используется интеллектуальный метод под названием «Метод линейной регрессии». Данный метод позволяет построить регрессионную модель, в которой можем увидеть зависимость изменения качества воздуха, а также построить саму линию регрессии, которая покажет, что с данными будет происходить дальше, то есть какой будет вероятный их прогноз и в каком промежутке.

Метод линейной регрессии использует уравнение линейной регрессии $y = a + bx$ (это уравнение похоже на уравнение прямой), где y – зависимая переменная, a – свободный член (пересечение) линии оценки, b – угловой коэффициент или градиент оценённой линии, x – независимая переменная.

Теперь приступим к использованию самой программы «RStudio» и начнем использовать метод линейной регрессии на данные загрязнения воздуха в Лондоне. Для начала выявим чему равен R^2 (коэффициент детерминации) используя функцию «lm» для Пригорода и Центра города на наличие Диоксид азота (NO₂) и оксиды азота (NO)

```
Call:
lm(formula = London.Mean.Roadside.Nitric.Oxide..ug.m3. ~ London.Mean.Background.Nitric.Oxide..ug.m3.,
    data = a)

Residuals:
    Min       1Q   Median       3Q      Max
-28.557  -9.409   3.265   7.502  19.724

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    43.03916    2.15591   19.96  <2e-16 ***
London.Mean.Background.Nitric.Oxide..ug.m3.  1.68437    0.07874   21.39  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.14 on 96 degrees of freedom
(24 observations deleted due to missingness)
Multiple R-squared:  0.8266,    Adjusted R-squared:  0.8248
F-statistic: 457.6 on 1 and 96 DF,  p-value: < 2.2e-16
```

Рис.1. Результат функции «lm»

Как видно на рисунке 1 изображен результат функции lm в которой виден R^2 равный 0,82 это значит, что коэффициент R^2 больше, чем 0,8 что регрессия имеет хорошую связь.

Ниже приведены данные по примесям в двух районах Лондона в пригороде и центре города в воздухе:

1. Диоксид азота (NO₂) равен 0.8032.

2. Оксиды азота (NO) равен 0.8483.
3. Озон (O₃) равен 0.9502.
4. PM10 равен 0.9425.
5. Диоксид серы равен 0.1111.

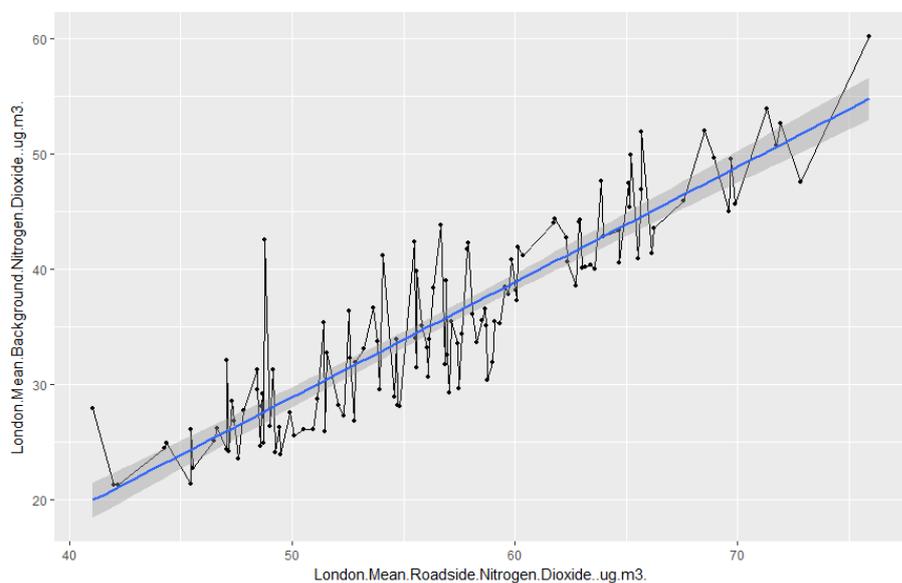


Рис.2 График зависимости диоксида азота в пригороде (Ox) и в центре города (Oy)

На рисунке 2 изображен график зависимости диоксида азота в пригороде и в центре города. На данном графике видно зависимости диоксид азота в Пригороде и в центре города. Данная регрессия показывает что прирост Диоксида азота в Пригороде Лондона растет быстрее чем в центре. Что показано на графике.

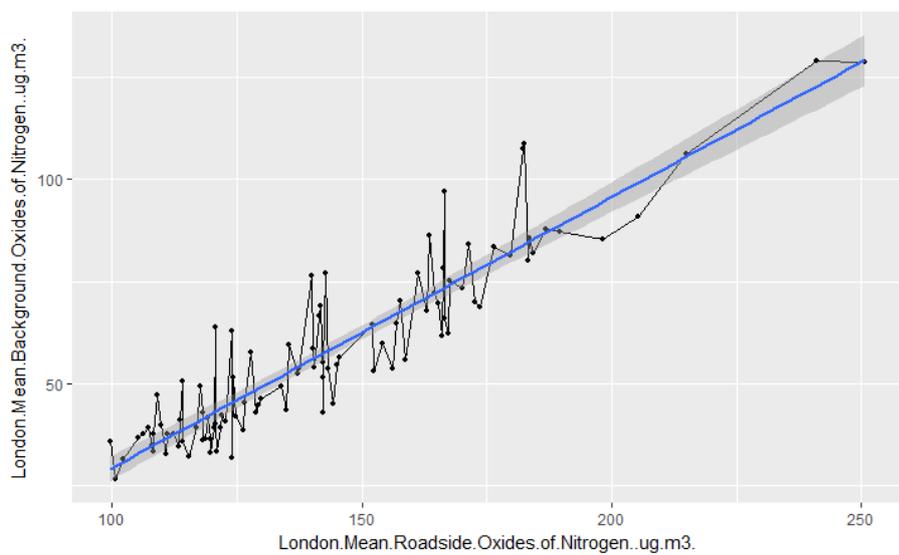


Рис. 3 график зависимости Оксидов азота в Пригороде (Ox) и Центре (Ox)

На рисунке 3 изображен график зависимости Оксид азота в Пригороде и Центре. На данном графике видно, что происходило с Оксидом азота в регионах и его изменения со временем постепенно растут.

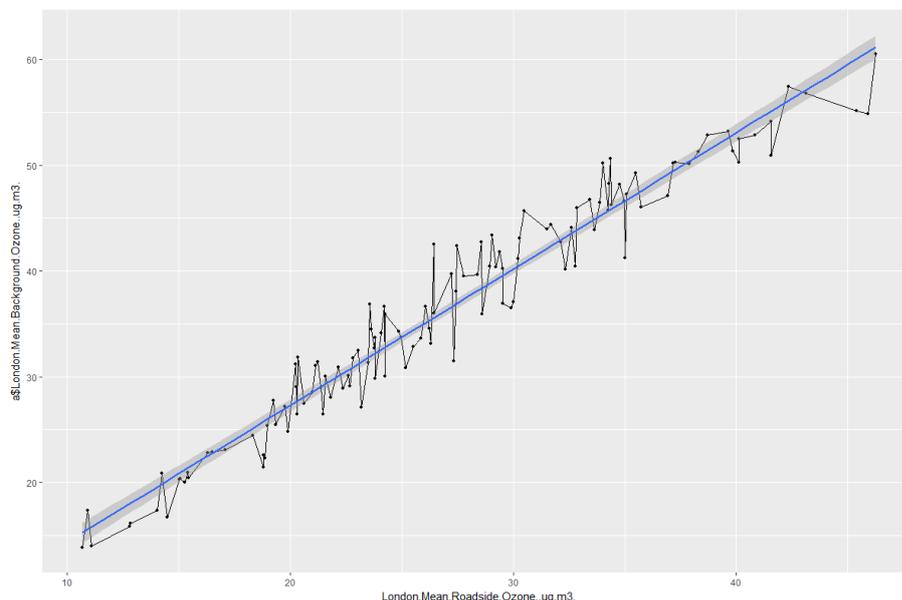


Рис. 4 график зависимости Озона (O3) в Пригороде (Ox) и Центре (Ox)

На рисунке изображен график зависимости Озона (O3) в Пригороде и Центре. На графике видно, как изменяется прирост Озона в разных регионах, а также видно, что график имеет очень сильную кривизну и ломаных прямых и это значит, что именно эта примесь очень сильно изменчива. А также в конце видно, что в Центре Лондона его относительно меньше, чем в Пригороде.

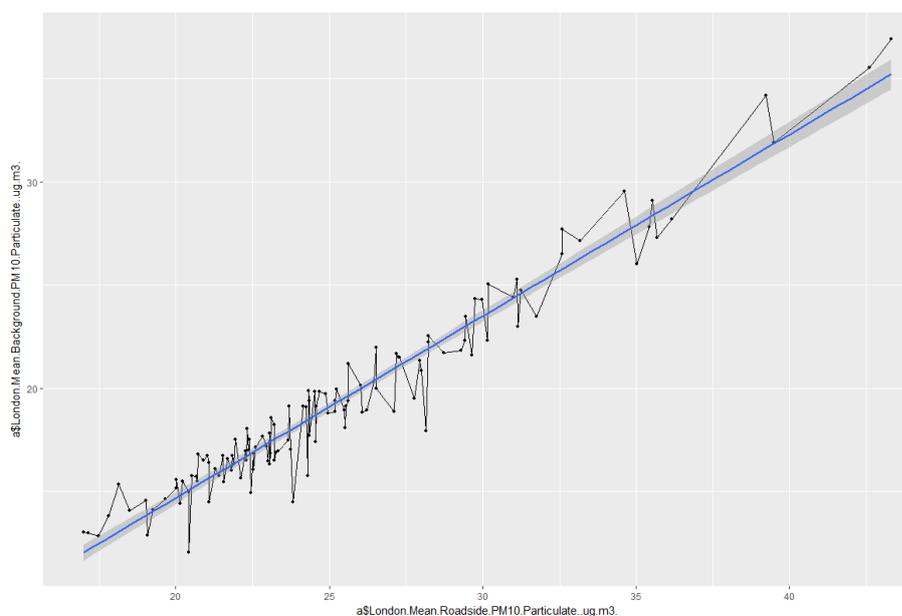


Рис. 5 график зависимости PM10 в Пригороде (Ox) и Центре (Oy)

На рисунке 5 изображен график зависимости PM10 в Пригороде и Центре. По графику видно, что он очень сильно не сбалансирован, , что в нем много ломаных прямых — это выражено тем, что данная примесь часто встречается в небольших количествах в воздухе.

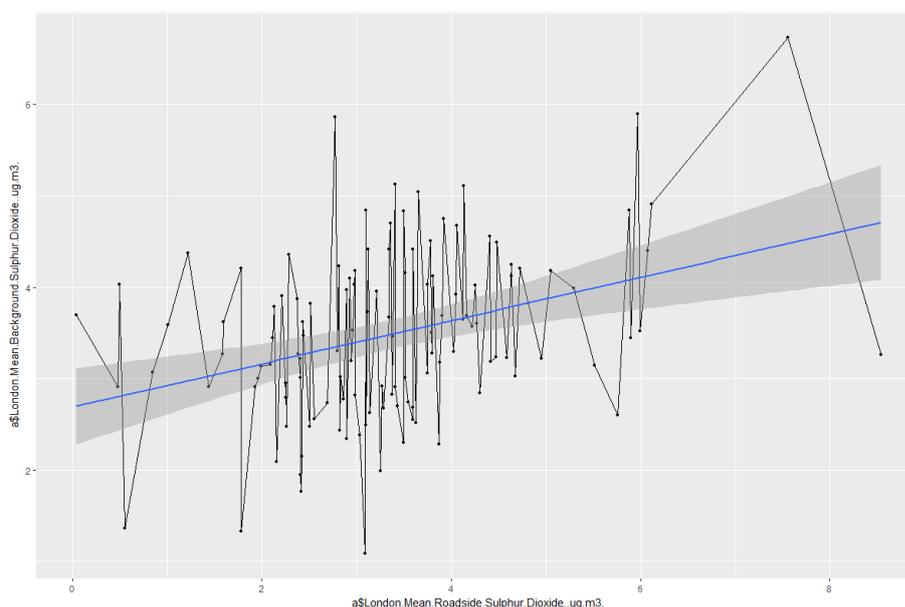


Рис. 6 график зависимости Диоксида серы в Пригороде (Ox) и Центре(Oy).

По графику можно сказать одно что у него очень плохая корреляция и это видно на графике, и она равна 0,1111, то есть очень слабая, потому что она меньше, чем 0.5 и практически нет прямой связи между диоксидом серы в пригороде и диоксидом серы в центре.

Таким образом, в данной статье была достигнута цель с выявлением зависимостей данных о загрязнении воздуха в двух регионах Лондона использованием метода линейной регрессии в системе R. Можно, судя по графикам, заметить, что в Пригороде Лондона примесей в воздухе не сильно мало, но меньше, чем в Центральной части. Это все можно заметить по графикам, которые были представлены выше.

Библиографический список

1. Стрижов В.В., Сологуб Р.А. Алгоритм выбора нелинейных регрессионных моделей с анализом гиперпараметров // Математические методы распознавания образов. 2009. Т. 14. № 1. С. 184-187.
2. Глухих И.Ю. Разработка моделей экспресс-анализа финансовой состоятельности организаций на базе методов многомерного регрессионного анализа // Управленческое консультирование. Актуальные проблемы государственного и муниципального управления. 2011. № 3 (43). С. 185-195.
3. Бугаевский Л.М., Прохоров Г.Г. Разработка методики составления карт взаимосвязи с использованием корреляционного и регрессионного анализов // Известия высших учебных заведений. Геодезия и

- аэрофотосъемка. 1990. № 6. С. 101-109.
4. Баджанов В.С., Матушевская Е.А. Применение корреляционно-регрессионного анализа для анализа себестоимости продукции на примере ГУП АО "Севастопольский Винодельческий завод" // Southern Almanac of Scientific Research. 2017. № 4 (4). С. 20-25.
 5. Манцаева А.А. Анализ долговременных тенденций производительности труда в РК: корреляционно-регрессионный анализ // Вестник Института комплексных исследований аридных территорий. 2015. Т. 1. № 1 (30). С. 18-24.
 6. Легкодух О.Ю., Капустина Д.Д., Кокодей Т.А. Анализ и прогноз динамики курса доллара, используя инструментарий регрессионного анализа // В сборнике: Развитие методологии современной экономической науки и менеджмента материалы I Всероссийской конференции студентов, аспирантов и молодых учёных. Севастопольский государственный университет. 2016. С. 57-58.
 7. Рашитова Н.Х. Анализ эффективности структуры экономики на основе корреляционно-регрессионного анализа // В сборнике: Инновационное развитие российской экономики материалы X Международной научно-практической конференции. Российской Федерации Российский экономический университет имени Г. В. Плеханова; Российский фонд фундаментальных исследований. 2017. С. 250-252.
 8. Boukezzoula R., Galichet S., Coquin D. From fuzzy regression to gradual regression: Interval-based analysis and extensions //Information Sciences. 2018. Т. 441. С. 18-40.
 9. de França F. O. A greedy search tree heuristic for symbolic regression //Information Sciences. 2018. Т. 442. С. 18-32.
 10. Georgiev I. et al. Testing for parameter instability in predictive regression models //Journal of Econometrics. 2018. Т. 204. №. 1. С. 101-118.
 11. База данных в Лондоне URL: <https://data.london.gov.uk/dataset> (дата обращения 07.05.2018)
 12. Основы линейной регрессии URL: <http://statistica.ru/theory/osnovy-lineynoy-regressii> (дата обращения 07.05.2018)
 13. Пример нахождения коэффициента детерминации URL: <https://math.semestr.ru/corel/prim2.php> (дата обращения 07.05.2018)
 14. Кочитов М.Е., Баженов Р.И. Исследование зависимости рабочих мест в Лондоне и Великобритании с использованием интеллектуального метода линейной регрессии в системе R // Постулат. 2018. №4
 15. London data store URL: <http://www.londonair.org.uk> (дата обращения: 09.04.2018)