МНОЖЕСТВЕНА РЕГРЕСИЯ ЗА ОЦЕНКА НА ВЛИЯНИЕТО НА НЯКОИ ФАКТОРИ ВЪРХУ СЕВЕРО-АТЛАНТИЧЕСКАТА ОСЦИЛАЦИЯ

Цветелина Величкова

Институт за изследване на климата, атмосферата и водите – Българска академия на науките
e-mail: tsvetelinavelichkova@abv.bg

Ключови думи: *Множествена регресия, Северо-Атлантическа осцилация, статистика*

Резюме: *Целта на настоящия доклад е да се оцени влиянието на някои фактори (озонът в двата центъра на действие на Северо-Атлантическата осцилация (NAO), въглеродният диоксид, броят слънчеви петна и вековите вариации на интензитета на геомагнитното поле) върху изменчивостта на NAO модата. Оценката е направена с помощта на множествената регресия. Анализът показа, че съществен принос в климатичните промени на Северноатлантическия регион в разглеждания период 1900 – 2019 г. имат озонът в ниската стратосфера и земното магнитно поле в Рейкявик и Понта Делгада.*

AN ESTIMATION OF SOME FACTORS INFLUENCING NORTH ATLANTIC OSCILLATION USING MULTIPLE REGRESSION

Tsvetelina Velichkova

Climate, atmosphere and water research institute – Bulgarian Academy of Sciences
e-mail: tsvetelinavelichkova@abv.bg

Keywords: *Multiple regression, North Atlantic Oscillation, statistics*

Abstract: *The purpose of this article is to estimate the influence of some factors (ozone in the two action centers of the North Atlantic Oscillation (NAO), the carbon dioxide, the number of sunspots and the centennial variations of geomagnetic field intensity) on the variability of the NAO mode. The evaluation was made using a multiple regression. The analysis showed that the significant contributions to climate change of the North Atlantic region, for the studied period 1900 – 2019, are the lower stratospheric ozone and geomagnetic field in Reykjavik and Ponta Delgada.*

Въведение

Основната цел на множествената регресия е да се състави модел, с който да се анализира влиянието на две или повече независими променливи върху една зависима променлива. Този метод позволява да се оцени както индивидуалното, така и общото влияние на тези фактори.

Функцията изразяваща връзката между X_i (независимите променливи) и Y (зависимата величина) се представя със следното уравнение:

(1) $$Y = a + b_1x_1 + b_2x_2 \dots b_nx_n + \varepsilon$$

тук *a* е свободният член на функцията, $b_1, b_2, \dots b_n$ се наричат регресионни коефициенти и показват връзката между *Y* и този *X*, пред който се намира съответният коефициент, ε е случайната грешка.

Независимите променливи, включени в модела, трябва да отговарят на следните изисквания:

1. Те трябва да са количествено измерими. Ако е наложително в модела да се включва качествен факторен показател, трябва да му се придаде количествено измерване, т.е. той трябва да се квантифицира.

2. Факторите да не са корелирани помежду си, т.е. да не са обвързани помежду си с функционална зависимост.

Неизпълнението на 2-то условие поражда проблема за колинеарност или мултиколинеарност на независимите променливи (в дясната страна на регресионното уравнение). Това води до завишаване оценките на дисперсиите (общата: $\sum(y-\bar{y})^2$; факторната: $\sum(\hat{y}-\bar{y})^2$ и остатъчната: $\sum(y-\hat{y})^2$), влошавайки параметрите на модела. Заключенията, относно влиянието на факторните променливи върху резултативния показател на модела, стават ненадеждни или некоректни. Проблемът мултиколинеарност води и до силна неустойчивост на оценките и грешките на модела (дори при несъществени изменения в изходните данни), до затруднения при определянето на самостоятелното влияние на отделните фактори и пр.

Адекватността на многофакторния регресионен модел се оценява посредством регресионния коефициент R (виж формула (2) и обясненията към нея), а също така и чрез F-критерия на Фишер, дефиниран с уравнението (3):

(2) $$\boldsymbol{R} = \sqrt{\left(\frac{\sum(\boldsymbol{Y_i}-\overline{\boldsymbol{Y}})^2-\sum(\boldsymbol{Y_i}-\widehat{\boldsymbol{Y_i}})^2}{\sum(\boldsymbol{Y_i}-\overline{\boldsymbol{Y}})^2}\right)}$$

Първият член в числителя на формула 2, както и знаменателят, означават пълната изменчивост на зависимата променлива Y, а вторият член в числителя – моделната грешка. Поради тази причина, регресионният коефициент винаги е положително число. Този критерий дава оценка на адекватността на модела дори и в случаите, когато зависимата променлива не е „нормално“ разпределена.

Прието е, че квадратът на регресионния коефициент R^2, умножен на 100, отразява процента на изменчивост на моделираната променлива, който регресионният модел е в състояние да опише.

(3) $$F = \frac{\frac{\sum(\hat{y}-\bar{y})^2}{m}}{\frac{\sum(y-\bar{y})^2}{n-m-1}} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}\cdot\frac{(n-m-1)}{m}$$

където числителят в (3) описва дисперсията на модела, а знаменателят – дисперсията на изходните данни; ***m*** са степените на свобода на многофакторния регресионен модел, а *n-m-1* са степените на свобода на остатъчната дисперсия.

Освен *общия* F-критерий, могат да се изчисляват и *частни* F-критерии, с които се оценява значимостта на всеки, включен в модела, фактор. Формулата за изчисляване на *частния* F-критерий, записана за фактора x_i има вида:

(4) $$F_{x_i} = \frac{R^2_{yx_1\dots x_i\dots x_m}-R^2_{yx_1\dots x_{i-1}\,x_{i+1}\dots x_m}}{1-R^2_{yx_1\dots x_i\dots x_m}}\cdot\frac{n-m-1}{1}$$

където $R^2_{x_1\dots x_i\dots x_m}$ е коефициентът на детерминираност на модела с включване на всички фактори; $R^2_{x_1\dots x_{i-1}\,x_{i+1}\dots x_m}$ е коефициентът на детерминираност на модела с изключен фактора x_i; *n* е броят на изходните данни използвани при статистическото моделиране, *m* – броят на независимите фактори включени в модела.

В следващите параграфи ще покажем резултатите получени с използването на множествената регресия.

Данни

Настоящото изследване включва анализ на сезонните данни за относителното съдържание на O_3 на 70 hPa за, взети от реанализа ERA 20 век, за месеците декември – март, за периода 1900-2019 г. Данните са взети в точките с приблизителните координати на Рейкявик, Исландия (60°с.ш.; 20° з.д.) и Понта Делгада, Азорските острови (40°с.ш.; 30°з.д) и са изгладени по 11 години. За анализа на въглеродния диоксид (CO_2) взехме неговите средно годишни стойности от обсерваторията Мауна Лоа в Хавай, както и исторически записи, получени от сондажни ядки на ледници в Антарктида. В изследването използвахме средно-годишни данни за броя на слънчевите петна (SSN) за периода 1900 – 2020г., които взехме от WDC-SILCO,

Кралската обсерватория на Белгия в Брюксел. Редът беше изгладен с 11- и 22- годишно плъзгащо се осредняване. Годишните данни за вековите изменения на геомагнитното поле (F_c) взехме от International Geomagnetic Reference Field model, взети отново за двете станции, определящи NAO индекса. 5-годишно плъзгащо се осредняване беше използвано за изглаждане на двата времеви реда на магнитното поле.

Месечните данни за атмосферния индекс NAO – определян като разликата в баричното налягане над Азорските острови и Исландия - взехме от Climatic Research Unit, University of East Anglia (https://crudata.uea.ac.uk/cru/data/nao/). За анализа използвахме средно месечно на индекса за декември – март в периода 1900 – 2021 г. Редът беше изгладен по 22 години, тъй като ни интересува дългопериодичната изменчивост на модата.

За да извършим оценката на влиянието на променливите върху NAO модата ще използваме многофакторната линейна техника, заложена в статистическия пакет STATISTICA 8.

Резултати

Изследваните независими променливи за периода 1900–2019 г. са: озона (O_3) на 70 hPa; въглеродния диоксид (CO_2); вековите вариации на земно магнитно поле (F_c); броя на слънчевите петна (SSN). Атмосферният NAO индекс е избран за зависима променлива.

Широка практика в научните изследвания е създаването на многофакторни регресионни модели, в които приноса на всеки от потенциалните фактори влияещи върху изменчивостта на анализираната променлива се определя вътрешно в модела като парциален регресионен коефициент. Общата формула за изчисляването на частичните корелационни (парциални) коефициенти е следната [2]:

(5) $$r_{yx_i.x_1x_2\dots x_m} = \frac{r_{yx_i.x_2\dots x_{m-1}} - r_{yx_m.x_1x_2\dots x_{m-1}} . r_{x_ix_m.x_1x_2\dots x_{m-1}}}{\sqrt{\left(1-r^2_{yx_m.x_1x_2\dots x_{m-1}}\right).\left(1-r^2_{x_ix_m.x_1x_2\dots x_{m-1}}\right)}}$$

където:

$r_{yx_i.x_1x_2\dots x_m}$ е частичният корелационен коефициент, измерващ зависимостта между *y* и x_i при изключване на влиянието на $x_1, x_2 \dots x_m$;

$r_{yx_m.x_1x_2\dots x_{m-1}}$ е частичният корелационен коефициент, измерващ зависимостта между y и x_m при изключване на влиянието на $x_1, x_2 \dots x_m$;

$r_{x_ix_m.x_1x_2\dots x_{m-1}}$ е частичният корелационен коефициент, измерващ зависимостта между x_i и x_m при изключване на влиянието на y;

Частичните корелационни коефициенти, обаче, нямат самостоятелно значение. Най-често те се използват в етапа на формулирането на модела за изключването на факторите с несъществено влияние върху зависимата променлива.

Като илюстрация на възможностите на многофакторния регресионен анализ уточнихме регресионните коефициенти на уравнението:

(6) $$NAO_{22} = a_0 + a_1.smtO_{3R} + a_2.CO_2 + a_3.smtF_{c_R} + a_4.SSN_{22}$$

отчитащо съвместното влияние на четири фактора върху променливата NAO_{22}. С $smtO_{3R}$ сме означили изгладените по 11 точки зимни стойности на озона за Рейкявик; CO_2 - редът с годишните стойности на въглеродния диоксид; $smtF_{cR}$ са вековите вариации на геомагнитното поле за Рейкявик, изгладени по 5 точки, а SSN_{22} – годишните стойности на броя на слънчевите петна, изгладени с 22-точков прозорец. Резултатите са представени в Таблица 1.

Статистически значимият коефициент на множествената регресия има стойност R=0,94, а справката с коефициента на детерминираност (R^2) показва, че моделът е в състояние да обясни 88% от пълната изменчивост на NAO индекса в изследвания период 1900–2019 г. Всички корелационни параметри са статистически значими, освен този на броя на слънчевите петна SSN.

Таблица 1. Оценка на свързаността между факторните променливи, участващи в множествения регресионен модел за Рейкявик

Redundancy of Independent Variables; Dependent Variabe: 22year smt. of winter NAO R-square column contains R-square of respective variable with all other independent variables				
	Toleran.	**R-square**	**Partial - Cor.**	**Semipart - Cor.**
11 yr. smt. wintO_3 for Reykjavik	0,591942	0,408058	0,756524	0,388547
ann. CO_2	0,503705	0,496295	0,240220	0,083117
5 year smt. F_c for Reykjavik	0,357673	0,642327	-0,828738	-0,497378
22 yr. avr. of tot. SSN	0,408040	0,591960	0,081046	0,027311

Анализът на регресионните коефициенти между факторните променливи, както и коефициента им на толеранс (представени в Таблица 1) показва добра независимост между тях. Освен това, анализът на парциалните и полу-парциалните регресионни коефициенти показва, че те „обясняват“ един добър процент от изменчивостта на Северо-Атлантическата осцилация. Магнитното поле и озонът са факторите с най-голям дял в изменчивостта на индекса в изследвания период.

Експериментът с множествената регресия бе повторен и за втория активен център на NAO индекса – Понта Делгада. Регресионното уравнение има вида:

(7) $$NAO_{22} = a_0 + a_1.smtO_{3PD} + a_2.CO_2 + a_3.smtF_{c_{PD}} + a_4.SSN_{22}$$

където $smtO_{3PD}$ е зимният озон за Понта Делгада, изгладен по 11 години; CO_2 - годишните стойности на въглеродния диоксид; $smtF_{cPD}$ – 5-годишните изгладени векови вариации на геомагнитното поле за Понта Делгада, а SSN_{22} – годишните стойности на броя на слънчевите петна, изгладени по 22 години. Резултатите са обобщени в Таблица 2.

Таблица 2. Оценка на свързаността между факторните променливи, участващи в множествения регресионен модел за Понта Делгада. Стойностите в червен цвят показват, че са статистически значими.

Redundancy of Independent Variables; Dependent Variabe: 22year smt. of winter NAO R-square column contains R-square of respective variable with all other independent variables				
	Toleran.	**R-square**	**Partial - Cor.**	**Semipart - Cor.**
11 yr. smt. wintO_3 for Ponta Delgada	0,411241	0,588759	-0,644504	-0,367851
ann. CO_2	0,660069	0,339931	0,260467	0,117730
5 year smt. F_c for Ponta Delgada	0,353478	0,646522	-0,498706	-0,251085
22 yr. avr. of tot. SSN	0,435633	0,564367	0,001219	0,000532

Линейният множествен регресионен коефициент е R=0,90. Моделът обяснява 81% от пълната изменчивост на NAO модата за периода 1900 – 2019 г. Корелационните параметри на три от факторите, участващи в моделното уравнение, са статистически значими. Коефициентът на SSN отново е статистически незначим. Ако броят на слънчевите петна бъде изключен от регресионното уравнение, то регресионният коефициент си запазва стойността, тоест в този случай SSN не играе роля в крайния резултат. Коефициентът на толеранс и регресионните коефициенти между факторните променливи демонстрират добра независимост между тях, както и техните парциални и полу-парциални коефициенти. Както и за северния активен център на модата – Рейкявик, така и за Понта Делгада, най-голям принос имат озона и магнитното поле.

Заключение

Този експеримент с използването на множествената линейна регресия показва, че подбирането на факторните променливи в регресионното уравнение трябва да става много внимателно, с отчитането на потенциалните зависимости между тях. Ако такава предварителна информация не съществува, то използването на разнообразни методи за определянето на адекватността на регресионния модел е абсолютно наложителна.

Благодарности

Авторът изказва благодарност на Фонд Научни изследвания - договор № КП-06-М54/1 от 15.11.2021, с чиято помощ бе осъществено изследването.

Литература:

1. Usoskin, I. G., Mursula, K., Solanki, S. K., Schuessler, M., Kovaltsov, G. A., 2002, A physical reconstruction of cosmic ray intensity since 1610, J. Geophys. Res., 107(A11), doi:10.1029/2002JA009343.
2. Петков, П. Иконометрия с Gretl и Exel, изд. Стопанска Академия « Д.А. Ценов», Свищов, 2010, ISBN 979-954-23-0452-4, 474 стр.