Известия Гомельского государственного университета
имени Ф. Скорины, №5(50), Ч1, 2008

УДК 519.24

Методы множественного регрессионного анализа при прогнозировании процессов миграции радионуклидов из почвы в растения

С. П. Жогаль, С. И. Жогаль, Н. Б. Осипенко, Т. С. Запольская, Н. Н. Запольский

На основе использования методов регрессионного и корреляционного анализа получены модели, отражающие зависимость коэффициента перехода Cs-137 из почвы в растение от степени загрязнённости почвы, её физико-химического и механического состава.

1 Постановка задачи, используемые процедуры статистического анализа данных

Исследование зависимостей и взаимосвязей между объективно существующими процессами и явлениями играют в современных прикладных науках большую роль. Выявление количественных соотношений в виде регрессионных зависимостей и сравнение натурных данных с теми, которые получены путем подстановки в уравнение регрессии значений объясняющих переменных, дают возможность лучше понять природу исследуемых явлений.

В связи с необходимостью решения проблемы уменьшения последствий аварии на Чернобыльской атомной электростанции (ЧАЭС) возник целый ряд научных задач. Одной из таких задач является получение прогнозов по накоплению радионуклидов в растениях и возможности его уменьшения. Уровень накопления радионуклидов в растениях зависит от того, какое это растение, к какому виду оно относится, от степени загрязнения радионуклидами почвы, от ее физико-химического и механического состава. На территории Республики Беларусь существует достаточное число стационарных сельскохозяйственных площадок, на которых осуществляются наблюдения и собираются статистические данные. Задача проведения анализа полученных данных методами математической статистики в целях получения определенных регрессионных зависимостей является весьма актуальной. В данной работе предпринята попытка построения многомерной множественной регрессионной зависимости такого важного показателя как коэффициент перехода (КП) радионуклидов из почвы в растения от целого ряда факторов – объясняющих признаков, таких как степень загрязненности почвы, ее физико-химические и механические характеристики. Задача построения регрессионных зависимостей весьма актуальна в связи с необходимостью прогнозирования последствий радиационного загрязнения Республики Беларусь.

Все исходные данные по стационарным площадкам, на которых производились замеры, были введены нами в dbf-файл, имеющий следующую структуру:

1) NAS_P(тип-char16) – наименование населенного пункта;
2) XOZ (char18) – наименование хозяйства;
3) RAION(char16) – наименование района;
4) GOD(Num 2.0) – год проведения измерений (последние 2 цифры);
5) TKULT(Num 2.0) – тип культуры;
6) T_POCHV(Num 1.0) – тип почвы;
7) GS_POCH(Num 6.0) – содержание Cs-137 в почве (пКюри/кг);
8) SR_POCH(Num 4.0) – содержание Sr-90 в почве (пКюри/кг);
9) CAO (Num 4.2) – содержание CaO в почве (мг/кг);
10) MGO(Num 6.2) – содержание в почве MgO (мг/100);
11) GIDR_K(Num 3.2) – гидролитическая кислотность почвы (мг-экв/100г);

12) EMK_POGL(Num 5.2) – емкость поглощения (мгэкв/100);
13) PH(Num 5.2) – содержание PH в почве;
14) GUMUS(Num 4.2) – содержание гумуса в почве;
15) K2O(Num 4.1) – содержание K2O в почве (мг/100г);
16) P2O5(Num 4.1) – содержание в почве P2O5 (мг/100);
17) CS_RAST(Num5.0) – содержание в растении Cs-137 (пКюри/кг);
18) KAL(Num 4.1) – содержание калия в растении (г/кг);
19) SR_RAST (Num 4.0) – содержание в растении Sr-90 (пКюри/кг);
20) CA(Num 3.1) – содержание в растении кальция (г/кг);
21) KJ_CS(Num 4.2) – коэффициент перехода (КП) цезия-137 из почвы в растение;
22) KH_SR(Num 4.2) – коэффициент перехода (КП) стронция-90 из почвы в растение.

Поскольку лишь данные по кормовым угодьям и произрастающим на них многолетним травам обладали достаточной репрезентативностью, мы ограничились выборочными данными по ним.

Что касается шестого поля T_POCHV, то при внесении в него записей была использована следующая кодировка, несущая информацию не только о типе почвы, но и о ее механическом составе. В поле T_POCHV в качестве записи вносились цифры 1,...,9 в соответствии с типом почвы и ее механическим составом. Соответствие может быть представлено в виде таблицы 1.1.

К сожалению, отсутствие ряда важнейших характеристик, непосредственно связанных с формированием коэффициентов перехода, в особенности, физико-химических характеристик выпавших радиоактивных осадков и степени увлажненности почвы по годам, не позволяет провести достаточно полный анализ информативности признаков. Однако, опираясь только на имеющиеся исходные данные можно провести статистический анализ и построить регрессионную модель, отражающую зависимость КП от имеющихся в нашем распоряжении данных по объясняющим признакам.

Таблица 1.1 – Механический состав почвы

1	Дерново-подзолистые	супесчаные
2	Дерново-глеевые	суглинки
3	Дерново-подзолистые	суглинки
4	Дерново-глеевые	супесчаные
5	Пойменные дерновые	супесчаные
6	Пойменные дерновые глееватые	
7	Пойменные дерново-глеевые	
8	Пойменные торфяно-глеевые	суглинки

Для осуществления первичного статистического анализа данных и построения регрессионных зависимостей, в которых в качестве целевого признака y выступает соответствующий КП, а в качестве объясняющих признаков x_i, i=1,2,...,n (n – число объясняющих признаков) фигурируют физико-механические и агрохимические свойства почвы, были использованы возможности широко известного пакета STATGRAFICS.

Необходимость включения в процесс исследования статистических связей в системе “почва-растение” в качестве одного из важнейших этапов построения множественной линейной регрессии очевидна, поскольку на коэффициент перехода радионуклидов из почвы в растение влияет целый ряд факторов, связанных с физико-механическими и агрохимическими свойствами почвы.

В общем виде модель множественной линейной регрессии записывается следующим образом [1-4]:

$$Y=b_0+b_1x_1+b_2x_2+...+b_nx_n+e, \quad (1.1)$$

где b_0, b_1,..., b_n – неизвестные параметры, e – вектор случайных ошибок с нулевыми математическими ожиданиями и дисперсией.

Важное значение для определения меры линейной зависимости между двумя какими-либо переменными из y, x_1, x_2,..., x_n после "вычитания" эффекта, обусловленного взаимодействием этих двух переменных с некоторым непустым множеством из оставшихся n-1 переменных, имеет частный коэффициент корреляции [5, 6].

При исследовании статистических взаимосвязей в исследуемой системе целесообразно также применение метода анализа главных компонент, позволяющего перейти в задаче от объясняющих признаков к их линейным комбинациям, обладающим целым рядом свойств, облегчающих поиск статистических закономерностей в системе: построенные на основе объясняющих факторов главные компоненты не коррелируют друг с другом и упорядочены по степени их вклада в суммарную дисперсию системы [5, 7].

2 Результаты применения методов регрессионного анализа для выявления статистических связей в системе "почва-растение"

2.1 Проверка исходной выборки на однородность и выделение однородных подвыборок

Исходная выборка по кормовым угодьям была проанализирована нами по критерию зависимости КП от типа почвы и года, в котором проводились измерения. Были выделены четыре достаточно различающихся между собой однородные выборки К1, К2, К3, К4. Полученные подвыборки имеют отчетливое различие по КП. Данный факт подтверждается также и тем, что попарное сравнение этих выборок с целью проверки их на однородность по критерию Колмогорова-Смирнова показало, что соответствующие гипотезы были отвергнуты с уровнем значимости α=0.05. Критерии формирования подвыборок К1, К2, К3, К4 представлены в следующей таблице.

Таблица 2.1

Подвыборка	Вид почвы	Год измерения
К1	непойменные	1987-1989
К2	пойменные	1987-1989
К3	непойменные	1989-1992
К4	пойменные	1989-1992

2.2 Корреляционный анализ зависимостей признаков

Проведенный анализ данных показал, что говорить о целесообразности использования простых регрессионных зависимостей типа $y=b_0+b_1x$, где y – коэффициент перехода KH_CS, x – один из объясняющих признаков, не приходится, поскольку величина SR – доля объясняемой за счёт регрессионной модели дисперсии КП очень мала. Кроме того, вследствие взаимной коррелированности всех компонент системы делать окончательные выводы о степени влияния объясняющих признаков на КП на основе простых коэффициентов корреляции нецелесообразно.

Например, при построении множественной модели зависимости KH_CS от объясняющих признаков встаёт вопрос об исключении из модели наименее коррелируемых с КП признаков. На основании величин коэффициентов простой корреляции KH_CS с объясняющими признаками по данным для общей выборки KPKU можно было бы сделать вывод о том, что PH практически не коррелируется с KH_CS. Однако это не так: отбрасывание из регрессионной модели слагаемого с PH приводит к значительному уменьшению показателя SR модели (SR=0.1756 вместо SR=0.2070) и увеличению стандартной ошибки (SE=1.64 вместо SE=1.6075). Это объясняется тем, что в случае коррелированности признаков необходимо

опираться в выводах не на величины простых коэффициентов корреляции, а на величины "очищенных" частных коэффициентов корреляции.

2.3 Построение моделей множественной регрессии

Для общей выборки KPKU, а также для выборок К1, К2, К3, К4 были получены матрицы частных коэффициентов корреляции по всем объясняющим признакам и KH_CS. Построенные матрицы содержат частные коэффициентов корреляции, определяющие линейную взаимосвязь KH_CS, CAO, MGO, P2O5, K2O, T_POCHV, GOD,CS_POCH, GIDR_K, EMK_POGL, PH, GUMUS друг с другом за вычетом "эффекта" взаимодействия всех остальных признаков.

По всем пяти выборкам с использованием программ построения множественной линейной регрессионной зависимости пакета STATGRAFICS были получены уравнения $y=a_0+a_1x_1+a_2x_2+...+a_nx_n$, где в качестве выходной переменной y выступал коэффициент перехода KH_CS, а в качестве переменных x_1, x_2,..., x_n – объясняющие признаки CAO, CS_POCH, P2O5, GIDR_K, MGO, EMK_POGL, K2O, PH, T_POCH, GUMUS, GOD соответственно.

Однако построенные регрессионные модели не следует считать окончательными как с точки зрения их точности, так и с точки зрения числа объясняющих признаков. В данных моделях представлен весь перечень признаков, которые могут оказать влияние на КП. Однако среди них есть признаки, которые практически не коррелируют с КП, т.е. статистически незначимы. Их исключение из построенных моделей только улучшило степень адекватности последних. Отметим ещё раз, что при определении статистически незначимых объясняющих признаков мы руководствовались не простыми коэффициентами корреляции, а частными коэффициентами корреляции этих признаков с целевой переменной, т.к. при проведении процедуры отбрасывания статистически незначимых признаков это действительно приводит к получению множественной регрессионной модели, обладающей более лучшими прогностическими характеристиками.

Рассмотрим данный подход на примере выборки KPKU. Используя матрицу частных корреляций, при построении оптимизированной регрессионной модели мы действовали поэтапно, отбрасывая каждый раз наиболее незначимый из оставшихся признаков.

Шаг 1. Отбрасываем EMK_POGL, в результате для полученной регрессионной модели имеем:

SR=0.2137, SE=1.600757.

Исходя из полученных значений для SR, SE заключаем, что качество модели несколько улучшилось.

Шаг 2. Отбрасываем наименее коррелируемый с KH_CS из оставшихся признаков--K2O, получаем:

SR=0.2185, SE=1.595845 –

качество модели улучшилось.

Шаг 3. Отбрасываем MGO:

SR=0.2215, SE=1.59277 –

качество модели улучшилось.

Шаг 4. Отбрасываем CAO:

SR=0.22149, SE=1.59281 –

качество модели, хотя и незначительно, но ухудшилось.

Следовательно, оптимальной с точки зрения её прогностических качеств будет множественная регрессионная зависимость KH_CS от значимых объясняющих признаков GOD, CS_POCH, N_POCHV, PH, GIDR_K, GUMUS, P2O5, CAO (для всей выборки KPKU).

Проводя подобную процедуру для построения регрессионных зависимостей по всем выборкам, мы получили оптимальное с точки зрения прогностических качеств модели по сравнению с первоначальными. Ниже приведена таблица, описывающая построенные по данному методу модели.

Таблица 2.2 – Регрессионные модели по выборкам

объясняющие признаки	КОЭФФИЦИЕНТЫ a_i				
	K1	K2	K3	K4	KP_KU
CONSTANT	95.23	-52.851	24.321	59.248	20.337
CAO				0.0007	-0.001
P2O5	-0.089	-0.166			-0.022
T_POCHV	0.524	1.172	0.243		0.254
GOD	-1.048	0.450	-0.203	-0.702	-0.247
CS_POCH	-0.000064	-0.000015	-0.000057	-5.464	-9.348
GIDR_K		0.279	-0.388	0.243	0.157
PH		1.372	-0.646	1.084	0.599
GUMUS	-0.336				-0.226
MGO	-0.049			0.018	
K2O	0.059	0.038	0.024	0.055	
EMK_POGL		-0.032		0.038	
SR	0.520	0.190	0.317	0.272	0.221
SE	1.396	1.821	0.940	1.627	1.592

Следует отметить, что данным подходом построения регрессионных зависимостей можно воспользоваться и при прогнозировании не только КП, но и объясняющих признаков, а также при процедуре заполнения пропусков, вызванных отсутствием ряда измерений тех или иных признаков. В любом случае, конечно, необходимо, чтобы измеренные исходные данные были максимально корректны. Более того, желательно, чтобы измерения проводились по каждому признаку параллельно несколько раз и результаты их параллельных наблюдений поставлялись аналитикам без их предварительного усреднения. Это позволит проводить более научно-обоснованные исследования по выявлению статистических связей в системе, проводить анализ адекватности построенных моделей.

Нами были проведены также и исследования по построению моделей множественной регрессии на основе метода анализа главных компонет.

2.4 Применение метода анализа главных компонент

Метод анализа главных компонент применялся нами для получения комплексных некоррелированных между собой факторов, синтезирующих действие на KH_CS следующих объясняющих признаков: CAO, MGO, P2O5, K2O, GIDR_K, EMK_POGO, PH, GUMUS.

Например, по выборке KPKU с помощью программ STATGRAFICSa, реализующих метод главных компонент, на основе перечисленных восьми признаков были сформированы восемь факторов главных компонент PCOMP1,...,PCOMP8. Целесообразность применения главных компонент вытекает из того, что, во-первых, они некорреллированы между собой, а во-вторых, они упорядочены по вкладу в суммарную дисперсию объясняющих признаков, по которым они строятся.

Первый факт позволяет при построении множественных регрессионных моделей, заменяя исходные признаки на базирующиеся на них главные компоненты, определять модельную значимость главных компонент, не прибегая к вычислению частных коэффициентов корреляции их между собой (т.к. они некоррелированы) и с KH_CS, а непосредственно исходя из величин вычисленных простых коэффициентов корреляции между KH_CS и главными компонентами. Приведём таблицу, характеризующую коррелированность главных ком-

понент с KH_CS (по выборке KPKU).

Таблица 2.3

ГЛАВНАЯ КОМПОНЕНТА	КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ KH_CS	SE	SR
PCOMP1	-0.170	1.786	2.9 %
PCOMP2	-0.023	1.812	0.06 %
PCOMP3	0.036	1.811	0.13 %
PCOMP4	-0.013	1.812	0.02 %
PCOMP5	0.051	1.810	0.27 %
PCOMP6	0.006	1.812	0.0 %
PCOMP7	0.082	1.806	0.69 %
PCOMP8	-0.161	1.789	2.6 %

Как видно из приведенной таблицы, корреляции главных компонент с КП малы, но из этого не стоит делать выводов о качестве множественной регрессионной модели, построенной на их базе и с учетом таких объясняющих признаков, как GOD, T_POCHV, CS_POCH, которые не включались в группу признаков для построения главных компонент, т.к. представляют собой относительно самостоятельные факторы.

Применяя ранее описанную процедуру построения оптимальной множественной регрессии, мы уменьшили количество главных компонент, участвующих в построении регрессии, до двух за счет исключения незначимых с точки зрения корреляции с КП. Исключая поэтапно наименее значимые компоненты, мы в итоге получили наиболее оптимальную модель множественной регрессии, которая имеет вид:

KH_CS=24.541846+0.230487*T_POCHV-0.263965*GOD-0.00000934*CS_POCH-
0.189648*PCOMP1-0.680687*PCOMP8,
SR=0.2346, SE=1.579283.

Данная регрессионная зависимость значительно качественнее регрессии, построенной по всем объясняющим признакам. Более того, данное уравнение регрессии по своим прогностическим качествам превосходит оптимизированное уравнение регрессии, построение которого было описано нами ранее, несмотря на то, что последнее включает в качестве объясняющих признаков не пять, как в данном случае, а восемь компонент, и для которого SR=0.2215, SE=1.59277.

Следовательно, применение метода анализа главных компонент при проведении подобных исследований оправдано и может приводить к улучшению прогнозных качеств строящихся регрессионных моделей.

3. Заключение

В процессе анализа признаков в системе ”почва-растение” нами были получены следующие результаты.

На основе использования пакета STATGRAFICS и методов регрессионного и корреляционного анализа:

– выявлена существенная зависимость КП цезия-137 из почвы в растение от типа почвы и года проведения измерений, с учетом этого результата исходная общая выборка была разбита на четыре однородных подвыборки;

– был подтвержден на исследуемых данных тот факт (не всегда учитываемый исследователями), что при определении степени зависимости между факторами в многофакторных системах более информативным следует считать коэффициент частной корреляции, нежели коэффициент простой корреляции;

– построены множественные регрессионные модели для кормовых угодий с многолетними травами, описывающие зависимость КП Cs-137 из почвы в растение от степени за-

грязненности почвы и ее характеристик;

– с использованием метода пошагового регрессионного анализа была проведена оптимизация построенных регрессионных моделей;

– были построены регрессионные зависимости для КП Cs-137, в которых в качестве объясняющих признаков выступают главные компоненты, синтезированные на основе реальных объясняющих признаков, но не коррелирующие друг с другом.

Полученные результаты позволяют осуществлять качественный прогноз динамики КП радионуклидов из почвы в растение в зависимости:

– от времени, прошедшего после выпадения осадков;

– типа почвы;

– физико-химических составляющих почвы.

Они позволяют также установить те характеристики почвы и её состава, которые являются наиболее значимыми либо, наоборот, является незначимыми и их следует не принимать во внимание при решении рассматриваемой прикладной задачи.

Построенные регрессионные зависимости позволяют осуществлять прогноз динамики коэффициента перехода радионуклидов из почвы в растения в зависимости от степени загрязненности почвы, её типа и содержания в ней различных компонент, времени, прошедшего после выпадения радиоактивных осадков.

Результаты работы могут найти применение в различных разработках, связанных с созданием систем комплексного радиоэкологического мониторинга загрязнённых территорий.

Abstract. The paper presents statistical models constructed on the basis of application of the methods of multiple regression analysis which enable to forecast the dynamics of the coefficient of migrating radioactive nuclides from soil to plants due to the level of the soil contamination, its type and the content of various components in it, and due to the time that has passed since the radioactive fallouts.

Литература

1. Применение математических методов и ЭВМ. Планирование и обработка результатов эксперимента. — Мн.: Высшая школа, 1989.—218 с.

2. Жогаль, С.П. Основы регрессионного анализа и планирования эксперимента / С.П. Жогаль, С.И. Жогаль, И.В. Максимей; Гомель: ГГУ, 1997.—94 с.

3. Химмельблау, Д. Анализ процессов статистическими методами / Д. Химмельблау;– М.: Мир, 1973. – 958 с.

4. Айвазян, С.А. Прикладная статистика: Основы моделирования и первичная обработка данных / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин; М.: Финансы и статистика, 1983. – 471 с.

5. Афифи, А. Статистический анализ: подход с использованием ЭВМ / А. Афифи, С. Эйзен; М.: Мир, 1982. – 488с.

6. Ферстер, М. Методы корреляционного и регрессионного анализа / М. Ферстер, Б. Ренц; М.: Финансы и статистика, 1983. – 302 с.

7. Харман, Г.Г. Современный факторный анализ / Г.Г. Харман; М.: Статистика, 1972. – 403 с.

Гомельский государственный
университет имени Ф. Скорины

Поступило 10.05.08