УДК 681.31(031)

МЕТОДЫ ПРОГНОЗИРОВАНИЯ В ИНФОРМАЦИОННОЙ СИСТЕМЕ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА

Янаева Марина Викторовна
к.т.н., доцент

Мурлин Алексей Георгиевич
к.т.н., доцент

Мурлина Владислава Анатольевна
к.т.н., доцент
Кубанский государственный аграрный университет, Кубанский государственный технологический университет, г. Краснодар, Россия

Статья посвящена исследованию методик составления прогнозов в краткосрочный и долгосрочный периоды по количеству выбросов загрязняющих веществ и проведению поиска скрытых знаний в базе данных

Ключевые слова: ЭКОЛОГИЧЕСКИЙ МОНИТОРИНГ, МЕТОДЫ ПРОГНОЗИРОВАНИЯ, ИНФОРМАЦИОННАЯ СИСТЕМА

UDC 681.31(031)

ENVIRONMENTAL MONITORING SYSTEMS IN CONSTRUCTION ORGANIZATIONS

Yanaeva Marina Viktorovna
Cand.Tech.Sci, associate professor

Murlin Aleksey Georgievich
Cand.Tech.Sci, associate professor

Murlina Vladislava Anatolevna
Cand.Tech.Sci, associate professor
Kuban State Agrarian University, Kuban State Technological University, Krasnodar, Russia

The article deals with methods of forecasting in the short and long term by the number of emissions and conducting search for hidden knowledge in the database

Keywords: ENVIRONMENTAL MONITORING, FORECASTING METHODS, INFORMATION SYSTEM

Для создания любой автоматизированной системы необходим подготовительный этап, связанный с исследованием и описанием предметной области, объектов или процессов автоматизации, а так же различных видов взаимосвязей между ними. Под предметной областью будем понимать информацию о совокупности объектов автоматизации и их характеристиках, которая представляется в виде специальных структур данных, хранится в базе данных (БД) и используется пользователями для решения различных функциональных задач.

В качестве методик поиска скрытых зависимостей в предметной области рассмотрим корреляционный и регрессионный анализы.

Корреляционный анализ применяется для количественной оценки взаимосвязи двух наборов данных, представленных в безразмерном виде. Корреляционный анализ дает возможность установить, ассоциированы ли наборы данных по величине. Коэффициент корреляции, всегда

обозначаемый латинской буквой r, используется для определения наличия взаимосвязи между двумя свойствами.

Связь между признаками (по шкале Чеддока) может быть сильной, средней и слабой. Тесноту связи определяют по величине коэффициента корреляции, который может принимать значения от -1 до +1 включительно. Критерии оценки тесноты связи показаны на рисунке 1.

Величина коэффициента корреляции	0.1 - 0.3	0.3 - 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 1.0
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая
		средняя		сильная	

Рисунок 1– Количественные критерии оценки тесноты связи

Коэффициент корреляции Пирсона r, который является безразмерным индексом в интервале от -1,0 до 1,0 включительно, отражает степень линейной зависимости между двумя множествами данных.

Показатель тесноты связи между двумя признаками определяется по формуле линейного коэффициента корреляции:

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2] \cdot [n\sum y^2 - (\sum y)^2]}}, \quad (1)$$

где x - значение факторного признака, y - значение результативного признака, n - число пар данных.

Парная корреляция – это связь между двумя признаками: результативным и факторным или двумя факторными.

Варианты связи, характеризующие наличие или отсутствие линейной связи между признаками:

– большие значения из одного набора данных связаны с большими значениями другого набора (положительная корреляция) - наличие прямой линейной связи;

– малые значения одного набора связаны с большими значениями другого (отрицательная корреляция) - наличие отрицательной линейной связи;

– данные двух диапазонов никак не связаны (нулевая корреляция) - отсутствие линейной связи.

В качестве примера возьмем набор данных А. Необходимо определить наличие линейной связи между признаками x и y.

Для графического представления связи двух переменных использована система координат с осями, соответствующими переменным x и y. Построенный график, называемый диаграммой рассеивания, показан на рисунок 2. Данная диаграмма показывает, что низкие значения переменной x соответствуют низким значениям переменной y, высокие значения переменной x соответствуют высоким значениям переменной y. Этот пример демонстрирует наличие явной связи.

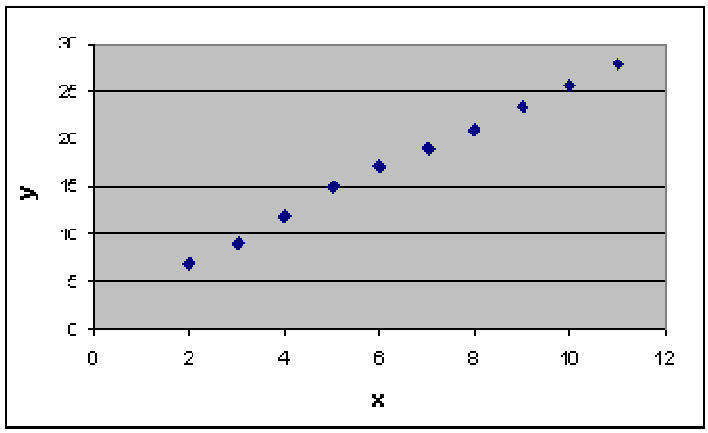

Рисунок 2– Диаграмма рассеивания

Таким образом, мы можем установить зависимость между переменными x и y. Рассчитаем коэффициент корреляции Пирсона между двумя массивами (x и y). В результате получаем значение коэффициент корреляции равный 0,998364, т.е. связь между переменными x и y является

весьма высокой. Любая зависимость между переменными обладает двумя важными свойствами: величиной и надежностью. Чем сильнее зависимость между двумя переменными, тем больше величина зависимости и тем легче предсказать значение одной переменной по значению другой переменной. Величину зависимости легче измерить, чем надежность. Надежность зависимости не менее важна, чем ее величина. Это свойство связано с представительностью исследуемой выборки. Надежность зависимости характеризует, насколько вероятно, что эта зависимость будет снова найдена на других данных.С ростом величины зависимости переменных ее надежность обычно возрастает.

Основная особенность регрессионного анализа: при его помощи можно получить конкретные сведения о том, какую форму и характер имеет зависимость между исследуемыми переменными.

Рассмотрим кратко этапы регрессионного анализа.

1. Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.

2.Определение зависимых и независимых (объясняющих) переменных.

3. Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель.

4. Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная).

5. Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии)

6. Оценка точности регрессионного анализа.

7. Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов.

8. Предсказание неизвестных значений зависимой переменной.

При помощи регрессионного анализа возможно решение задачи прогнозирования и классификации. Прогнозные значения вычисляются путем подстановки в уравнение регрессии параметров значений объясняющих переменных. Решение задачи классификации осуществляется таким образом: линия регрессии делит все множество объектов на два класса, и та часть множества, где значение функции больше нуля, принадлежит к одному классу, а та, где оно меньше нуля, к другому классу.

Рассмотрим основные задачи регрессионного анализа: установление формы зависимости, определение функции регрессии, оценка неизвестных значений зависимой переменной.

1. Установление формы зависимости.

Характер и форма зависимости между переменными могут образовывать следующие разновидности регрессии:

– положительная линейная регрессия (выражается в равномерном росте функции);

– положительная равноускорено возрастающая регрессия;

– положительная равнозамедленно возрастающая регрессия;

– отрицательная линейная регрессия (выражается в равномерном падении функции);

– отрицательная равноускорено убывающая регрессия;

– отрицательная равнозамедленно убывающая регрессия.

Однако описанные разновидности обычно встречаются не в чистом виде, а в сочетании друг с другом. В таком случае говорят о комбинированных формах регрессии.

Вторая задача сводится к выяснению действия на зависимую переменную главных факторов или причин, при неизменных прочих равных условиях, и при условии исключения воздействия на зависимую

переменную случайных элементов. Функция регрессии определяется в виде математического уравнения того или иного типа.

3. Оценка неизвестных значений зависимой переменной.

Решение этой задачи сводится к решению задачи одного из типов:

1. Оценка значений зависимой переменной внутри рассматриваемого интервала исходных данных, т.е. пропущенных значений; при этом решается задача интерполяции.

2. Оценка будущих значений зависимой переменной, т.е. нахождение значений вне заданного интервала исходных данных; при этом решается задача экстраполяции.

3. Обе задачи решаются путем подстановки в уравнение регрессии найденных оценок параметров значений независимых переменных. Результат решения уравнения представляет собой оценку значения целевой (зависимой) переменной.

4. Рассмотрим некоторые предположения, на которые опирается регрессионный анализ.

5. Предположение линейности, т.е. предполагается, что связь между рассматриваемыми переменными является линейной. Так, в рассматриваемом примере мы построили диаграмму рассеивания и смогли увидеть явную линейную связь. Если же на диаграмме рассеивания переменных мы видим явное отсутствие линейной связи, т.е. присутствует нелинейная связь, следует использовать нелинейные методы анализа.

6. Предположение о нормальности остатков. Оно допускает, что распределение разницы предсказанных и наблюдаемых значений является нормальным. Для визуального определения характера распределения можно воспользоваться гистограммами остатков.

7. При использовании регрессионного анализа следует учитывать его основное ограничение. Оно состоит в том, что регрессионный анализ

позволяет обнаружить лишь зависимости, а не связи, лежащие в основе этих зависимостей.

Регрессионный анализ дает возможность оценить степень связи между переменными путем вычисления предполагаемого значения переменной на основании нескольких известных значений. Уравнение регрессии выглядит следующим образом: Y=a+b*X. При помощи этого уравнения переменная Y выражается через константу a и угол наклона прямой (или угловой коэффициент) b, умноженный на значение переменной X. Константу a также называют свободным членом, а угловой коэффициент - коэффициентом регрессии или B-коэффициентом. В большинстве случав (если не всегда) наблюдается определенный разброс наблюдений относительно регрессионной прямой. Остаток - это отклонение отдельной точки (наблюдения) от линии регрессии (предсказанного значения). Входной интервал Y - это диапазон зависимых анализируемых данных, он должен включать один столбец. Если функция регрессии определена, интерпретирована и обоснована, и оценка точности регрессионного анализа соответствует требованиям, можно считать, что построенная модель и прогнозные значения обладают достаточной надежностью. Прогнозные значения, полученные таким способом, являются средними значениями, которые можно ожидать.

Возможно применение и интерполяции. Интерполяция – отыскание промежуточных значений величины по некоторым известным её значениям. Например, отыскание значений функции f (x) в точках x, лежащих между точками (узлами) x0 < x1 < ... <xn, по известным значениям yi = f (xi) (где i = 0, 1, ..., n). В случае, если x лежит вне интервала, заключённого между x0 и xn, аналогичная задача называется задачей экстраполяции. При простейшей линейной интерполяции значение f (x) в точке x, удовлетворяющей неравенствам x0 < x < x1, принимают равным значению линейной функции, совпадающей с f (x) в точках x = x0

и x = x1. Задача интерполяции, со строго математической точки зрения, является неопределённой, если про функцию f (x) ничего неизвестно, кроме её значений в точках x0, x1,..., xn, то её значение в точке x, отличной от всех этих точек, остаётся совершенно произвольным. Задача интерполяции приобретает определённый смысл, если функция f (x) и её производные подчинены некоторым неравенствам. Если, например, заданы значения f (x0) и f (x1) и известно, что при x0 < x < x1 выполняется неравенство |f''(x)| ≤ M, то погрешность формулы может быть оценена при помощи неравенства.

$$\left|f(x)-y\right| \le \frac{M}{2}(x-x_0)(x-x_1),$$

(2)

Рассмотрим еще один термин. Экстраполяция (от экстра... и лат. polio — приглаживаю, выправляю, изменяю) в математике и статистике, приближённое определение значений функции f (x) в точках x, лежащих вне отрезка [x0, xn], по её значениям в точках x0 < x1 <... <xn. Наиболее распространённым видом экстраполяции. является параболическая экстраполяция., при которой в качестве значения f (x) в точке x берётся значение многочлена Pn (x) степени n, принимающего в n + 1 точке xi заданные значения yi = f (x). Для параболической экстраполяции пользуются интерполяционными формулами.Рассмотрим основные интерполяционные формулы.

1. Интерполяционная формула Лагранжа:

$$f(x) \approx P_n(x) = \sum_{k=0}^{n} y_k \frac{(x-x_0)(x-x_1)\ldots(x-x_{k-1})(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_0)(x_k-x_1)\ldots(x_k-x_{k-1})(x_k-x_{k+1})\ldots(x_k-x_n)},$$

(3)

Ошибка, совершенная при замене функции f (x) выражением Pn(x), не превышает по абсолютной величине:

$$M\frac{|(x-x_0)(x-x_1)\ldots(x-x_n)|}{(n+1)!},\qquad (4)$$

где M — максимум абсолютной величины (n + 1)-й производной f n+1(x) функции f (x) на отрезке [x0, xn].

2. Интерполяционная формула Ньютона. Если точки x0, x1, ..., xn расположены на равных расстояниях (xk = x0 + kh), многочлен Pn(x) можно записать так:

$$P_n(x_0+th)=y_0+\frac{t}{1!}\Delta y_0+\frac{t(t-1)}{2!}\Delta^2 y_0+\ldots+\frac{t(t-1)\cdots(t-n+1)}{n!}\Delta^n y_0,$$

(5)

где x0 + th = x, а Δk — разности k-го порядка: Δkyi = Δk — 1 yi +1 — Δk — 1yi).

Это так называемая формула Ньютона для интерполирования вперёд; название формулы указывает на то, что она содержит заданные значения y, соответствующие узлам интерполяции, находящимся только вправо от x0. Эта формула удобна при интерполировании функций для значений x, близких к x0. При интерполировании функций для значений x, близких к наибольшему узлу xn, употребляется сходная формула Ньютона для интерполирования назад. При интерполировании функций для значений x, близких к xk, формулу Ньютона целесообразно преобразовать, изменив начало отсчёта (см. ниже формулы Стирлинга и Бесселя).

Формулу Ньютона можно записать и для неравноотстоящих узлов, прибегая для этой цели к разделённым разностям (см. Конечных разностей исчисление). В отличие от формулы Лагранжа, где каждый член зависит от всех узлов интерполяции, любой k-й член формулы Ньютона зависит от первых (от начала отсчёта) узлов и добавление новых узлов вызывает лишь добавление новых членов формулы (в этом преимущество формулы Ньютона).

3. Интерполяционная формула Стирлинга:

$$P_n(x_0+th)=y_0+\frac{t}{1!}\Delta y_0+\frac{t(t-1)}{2!}\Delta^2 y_0+\ldots+\frac{t(t-1)\cdots(t-n+1)}{n!}\Delta^n y_0 \quad (6)$$

Применяется при интерполировании функций для значений x, близких к одному из средних узлов a; в этом случае естественно взять нечётное число узлов x—k, ..., x—1, x0, x1, ..., xn, считая a центральным узлом x0.

4. Интерполяционная формула Бесселя:

$$f(x_0+th)\approx \mu y_{1/2}+\frac{(t-1/2)}{1!}\delta y_{1/2}+\frac{t(t-1)}{2!}\mu\delta^2 y_{1/2}+\frac{t(t-1)(t-1/2)}{3!}\delta^3 y_{1/2}+$$
$$+\frac{t(t-1)(t+1)(t-2)}{4!}\mu\delta^4 y_{1/2}+\frac{t(t-1)(t+1)(t-2)(t-1/2)}{5!}\delta^5 y_{1/2}+\cdots+$$
$$+\frac{t(t-1)(t+1)\cdots(t-k)(t+k-1)(t-1/2)}{(2k+1)!}\delta^{2k+1} y_{1/2} \quad (7)$$

Применяется при интерполировании функций для значений x, близких середине a между двумя узлами; здесь естественно брать чётное число узлов x—k, ..., x—1, x0, x1,..., xk, xk + 1, и располагать их симметрично относительно a (x0 < a < x1).

Рассмотренные методы реализованы в информационной системе в пакете анализа данных.

При помощи «Пакета анализа», который доступен в главном меню информационной системы экологического мониторинга, пользователь имеет возможность осуществить поиск скрытых знаний в базе данных с помощью методов корреляционного и регрессионного анализов. На рисунке 3 приведено исследование зависимости количества выбросов загрязняющих веществ от времени года. При этом справа отображаются данные по корреляционному анализу, по которым можно оценить существует ли зависимость или нет. В частности такими критерием является коэффициент Пирсона. На графике представлено построение линейной и параболической регрессии. Справа отображаются:

– уравнение линейной регрессии;

– оценка дисперсии случайной ошибки;

– количество степеней свободы;

– стандартная ошибка регрессии;

– t-критерий;

– стандартная ошибка регрессионного коэффициента A;

– стандартная ошибка регрессионного коэффициента B;

– регрессионный коэффициент A;

– регрессионный коэффициент B;

– коэффициент детерминации;

– уравнение параболической регрессии.

На диаграмме отображено процентное соотношение выбросов по веществам, что позволяет оценить какое вещество оказывает наибольшее влияние на атмосферу (рисунок 3).

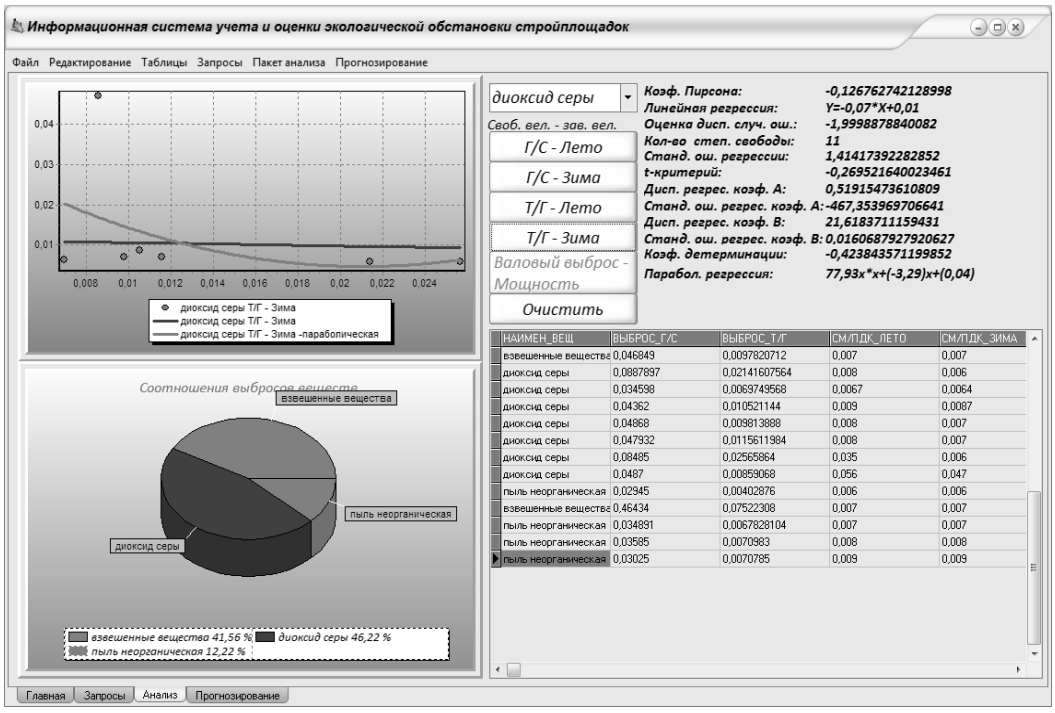

Рисунок 3 – Применение пакета анализа для поиска зависимостей

На рисунке 4 приведено исследование зависимости количества загрязняющего вещества от мощности двигателя.

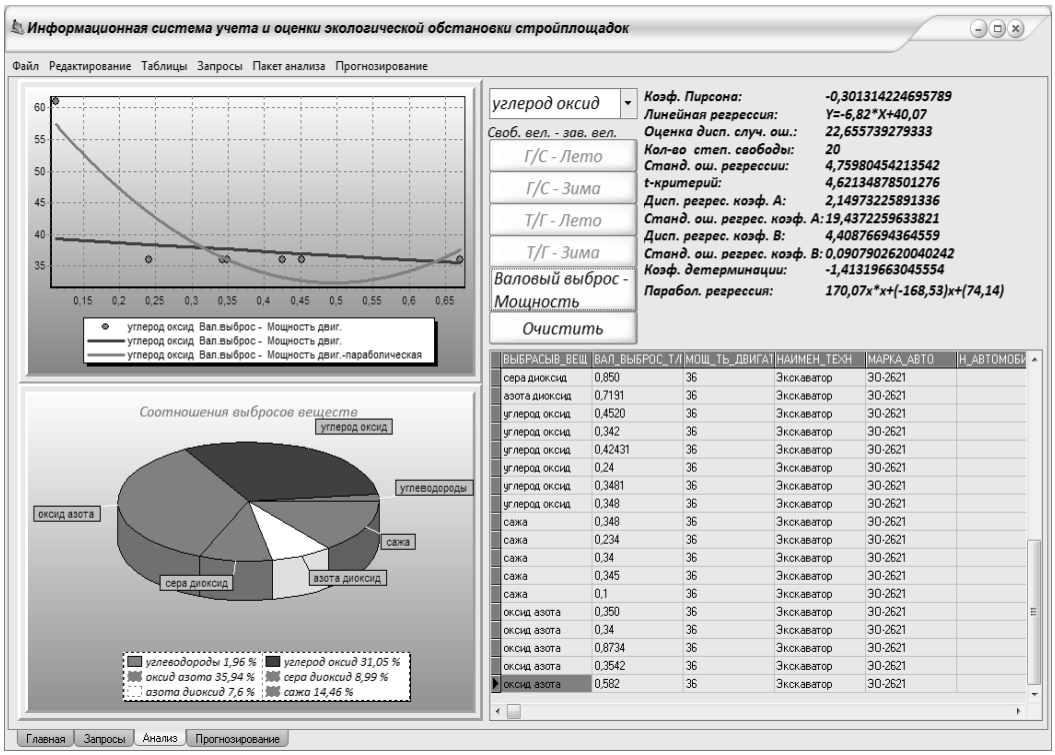

Рисунок 4 – Оценка зависимости количества выбросов от мощности двигателя

Также возможно на основе имеющих данных в базе данных спрогнозировать количество выбрасываемого вещества на определенный период. На рисунке 5 представлен прогноз рассеивания вещества «пыль неорганическая» на девятый месяц. При прогнозировании использован метод Лагранжа, приведенный на верхнем графике и первая интерполяционная формула Ньютона, график которой приведен на нижней части графика. Благодаря использованию двух методов пользователь имеет возможность оценить точность прогноза. На диаграмме приведено соотношение выбросов вещества по месяцам для наглядной оценки количества выбросов (рисунок 5).

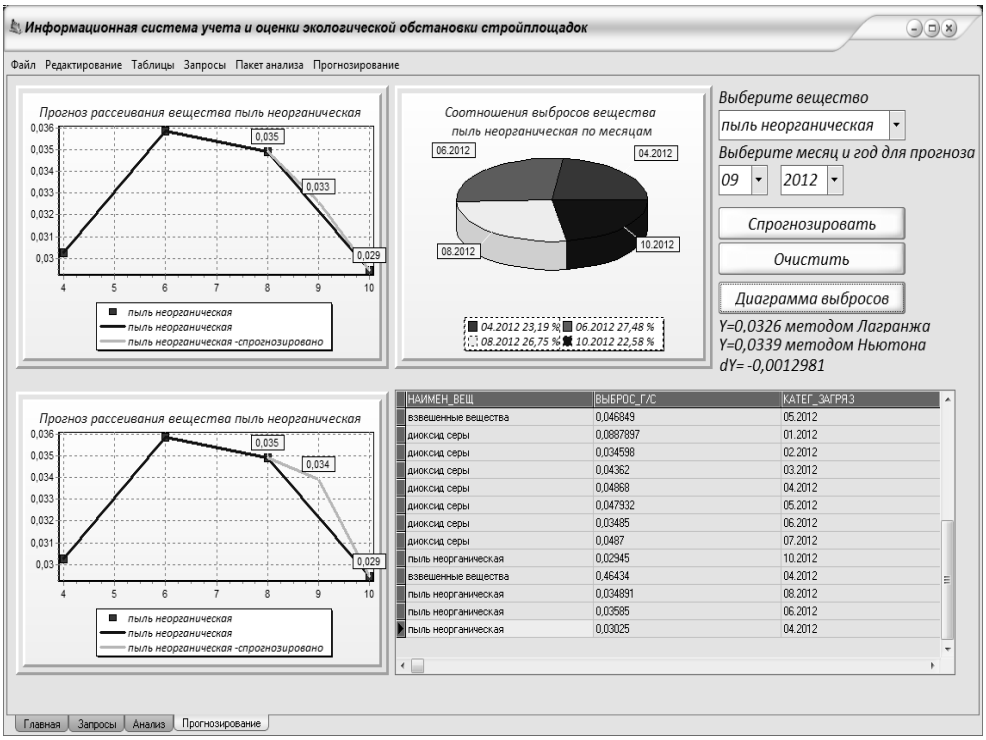

Рисунок 5 – Прогнозирование данных

Предложенные методики позволяют составлять прогнозы на будущее по количеству выбросов загрязняющих веществ и проводить поиск скрытых знаний в базе данных.

Список литературы

1. Цыгикало Т.И., Янаева М.В., Цыгикало Д.В., Руденко М.В., Автоматизация процесса управления экологическим мониторингом строительной площадки // Научный журнал КубГАУ [Электронный ресурс]. – Краснодар КубГАУ , 2012 . - №77. – шифр Информрегистра: 0421200012\0222. Режим доступа: http://ej.kubagro.ru/2012/03/pdf/70.pdf.

2. Том Кайт, Oracle для профессионалов, перевод с английского /Том Кайт – СПб.: ООО ДиаСофтЮП, 2003. - 672с.

3. Орлов С.А. Технологии разработки программного обеспечения. – СПб.: Питер, 2003. – 480с.