УДК 504.064

«АНАЛИЗ ПРИМЕНЕНИЯ СПОСОБОВ ЗАПОЛНЕНИЯ ПРОПУСКОВ В ДАННЫХ ВО ВРЕМЕННЫХ РЯДАХ В ЭКОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ»

Радчикова Е.С.

БГУ, географический факультет

Анализ временных рядов приобретает все большую популярность в самых разнообразных научных исследованиях. Широкое развитие он получил и в экологии, в связи с наличием большого количества динамических процессов, объемных массивов наблюдений с широким пространственным и временным диапазонами и невозможностью рассматривать многие явления без учета временного контекста. Особенно велико значение анализа временных рядов в экологическом прогнозировании, мониторинге состояния окружающей среды, моделировании различных процессов и явлений и множестве других экологических исследований. Использование анализа временных рядов напрямую связано со стремлением к достижению экологической безопасности и устойчивого развития.

Временной ряд отличается от простой выборки данных тем, что в нем данные упорядочены по времени. Предполагается, что данные содержат регулярную составляющую (одну или несколько) и случайную составляющую, которая затрудняет обнаружение регулярных компонент. Большинство регулярных составляющих временных рядов принадлежит к двум типам: они являются либо трендом, либо сезонной составляющей. Тренд представляет собой, как правило, монотонную компоненту, изменяющуюся во времени и описывающую общую тенденцию в изменении анализируемого признака. К трендам можно отнести такие явления, как увеличение среднегодовой температуры, сокращение площади экваториальных лесов, рост численности населения. Сезонная составляющая - это регулярная периодическая составляющая, описывающая периодически повторяющиеся колебания анализируемого признака. Например, это изменение температуры по сезонам года, многолетние циклы солнечной активности, высота воды в реке, в зависимости от ее годового режима. Оба эти вида регулярных компонент часто присутствуют во временном ряде одновременно, что необходимо учитывать при работе с временными рядами.[1]

Достаточно часто экологические данные содержат пропуски. Причины появления пропусков бывают различными: поломка приборов, неблагоприятные погодные условия, ошибки, повреждение носителей информации. Многие методы обработки временных рядов исключают возможность работать с неполными данными. Поэтому велика важность восстановления пропусков в таких данных.

Самым простым решением обработки данных, безусловно, является исключение некомплектных наблюдений, содержащих пропуски, и дальнейший анализ полученных таким образом «полных» данных. Однако понятно, что такой подход приводит к сильному различию статистических выводов, сделанных при наличии в данных пропусков и при их отсутствии. Поэтому более перспективным является иной путь – заполнение пропусков перед анализом фактических данных. Этот подход имеет явные преимущества: ясное представление структуры данных; вычисление необходимых итоговых значений; уверенная интерпретация результатов анализа, так как можно опираться на традиционные характеристики и суммарные значения. Выбор метода заполнения пропусков является непростой задачей, и зависит от различных факторов: наличия регулярных компонент и их особенностей, причин возникновения пропусков в данных и характера этих пропусков (случайный или нет), а также особенностей данных и проводимого исследования.[2,3]

В основных статистических пакетах (Statistica, SPSS) предусмотрены 5 стандартных методов заполнения пропусков: заполнение пропусков по средним значениям ряда, интерполяция по соседним точкам, среднее по N соседним точкам, медиана N ближайших значений ряда, заполнение пропусков прогнозами линейной регрессии. Также используются и другие методы – метод расчета среднего по соответствующей дате, метод поиска процента от максимума или другой знаковой величины.

В статье не рассматриваются более сложные методы, требующие дополнительного моделирования, а предлагаются самые простые, доступные и распространенные. Пропуски данных принимаются за случайные. Методы рассматриваются на примере среднемесячной температуры месяцев с мая по октябрь на глубине 0,5 метра в озере Баторино с 1966 по 1975 гг. На рисунке 1 представлен график изменения температуры без заполнения пропусков. Температурные данные по этому озеру имеются за период с 1955 по 2012 год, имеют четко выраженную сезонную составляющую и не имеют тренда. Для примера была взята одна из самых проблемных частей этих данных, в которой данные за 2 года практически отсутствуют.

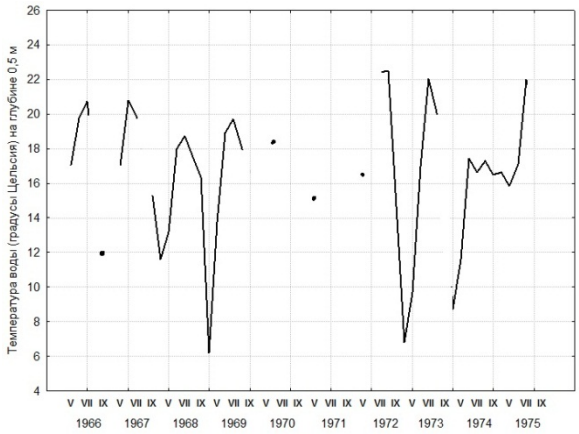

Рисунок 1. График изменения температуры воды на глубине 0,5 м

Метод среднего значения

При выборе этого метода все недостающие данные просто заменяются средним арифметическим значением для всех наблюдений. Данный метод может не подходить, когда ряд не постоянен или когда есть большие систематические колебания в переменных ряда. С другой стороны, полное среднее часто является лучшим априорным (беспристрастным) предположением для недостающих данных. Наиболее эффективен этот метод в работе с таким временным рядом, в котором проблематично выделить регулярную составляющую. Во временных рядах с сезонной составляющей либо трендом этот способ не имеет смысла. На рисунке 2 представлен график с заполнением пропусков средними значениями.

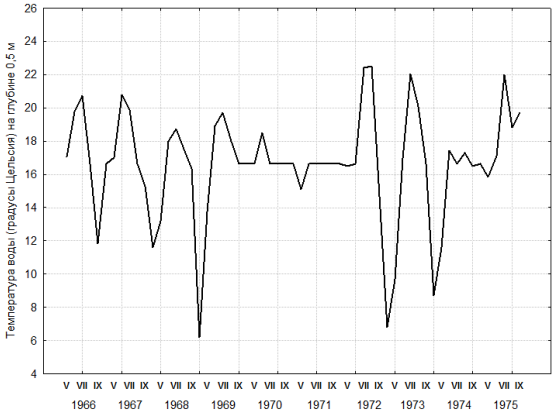

Рисунок 2. График изменения температуры с заполнением пропусков средними значениями

На рисунке видно, как нарушилось сезонное колебание при применении такого метода заполнения пропусков. Вместо общего среднего при наличии сезонной составляющей уместнее использовать среднее по соответствующей дате. На рисунке 3 пропуски заменены средним значением между присутствующими значениями температур соответствующего месяца.

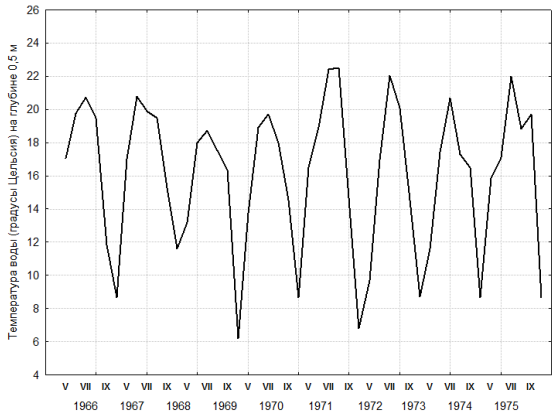

Рисунок 3. График изменения температуры с заполнением пропусков средними значениями по месяцам

Этот метод имеет много преимуществ, в сравнении с предыдущим, однако он не обладает гибкостью к несезонным изменениям и даже может привести к трудностям в определении линии тренда, сделать ее менее заметной для распознавания.

Метод интерполяции соседних точек

При выборе этого метода, недостающие данные вычисляются путем интерполяции из соседних не-недостающих точек. Графически этот метод сводится к замене недостающих данных путем соединения по прямой линии точки непосредственно перед отсутствующими данными с точкой сразу после отсутствующих данных (рисунок 4 для нашего примера). Этот метод предполагает, что есть некоторая серийная корреляция данных, то есть каждое наблюдение является в некоторой степени связаным с предыдущим и последующим, и, следовательно, наиболее похоже на предыдущее и следующее наблюдения. Метод линейной интерполяции достаточно точен только в том случае, когда значения переменной близки и изгиб кривой между ними невелик. Этот метод учитывает несезонные изменения, однако, в некоторых случаях он не может быть полезен. Например, температура самого теплого месяца, хоть и имеет некоторую связь с ближайшими измерениями, не является плотно связанной с ними, и ее значение выше прилегающих отметок.

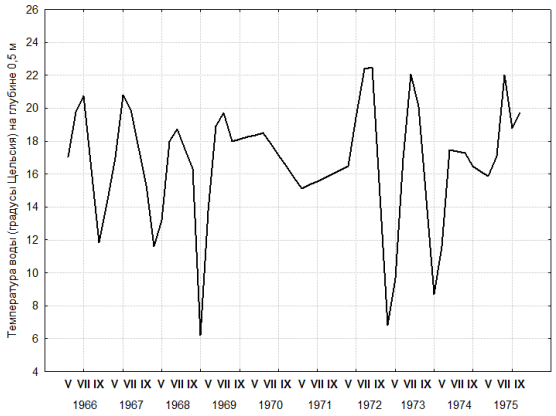

Рисунок 4. График изменения температуры с заполнением пропусков методом интерполяции

Среднее значение N соседних точек

При выборе этого метода, недостающие данные вычисляются из среднего значения N соседних точек по обе стороны пропуска. Например, когда N остается по умолчанию равным единице, то недостающие данные будут заменены средним между значением непосредственно перед пропуском и значением сразу после него. В целом, этот метод предполагает, что данные в области, заданной

параметром N, более похожи друг на друга, чем данные более удаленных точек. Этот метод имеет тот же недостаток, что и предыдущий – он не может быть применим при пропуске данных в точке максимума или минимума.

Медиана N соседних точек

Этот вариант заполнения отсутствующих данных по существу такой же, как предыдущий, за исключением того, что недостающие данные заменяются медианы N соседних точек. Медиана более точно характеризует данные с несимметричным распределением или с любым распределением, не соответствующим нормальному закону.

Заполнение прогнозными значениями линейной регрессии тренда

Еще один вариант, предлагаемый статистическими пакетами – метод наименьших квадратов линии регрессии для временных рядов. Отсутствующие данные заменяются значениями, предсказываемыми этой линией регрессии. Этот метод предполагает, что наиболее яркой особенностью серии является ее линейный тренд во времени. На примере (рисунок 5) результат заполнения получился схожим с заполнением средними значениями. Это связано с тем, что на исследуемых данных отсутствует тренд и велико значение сезонной составляющей.

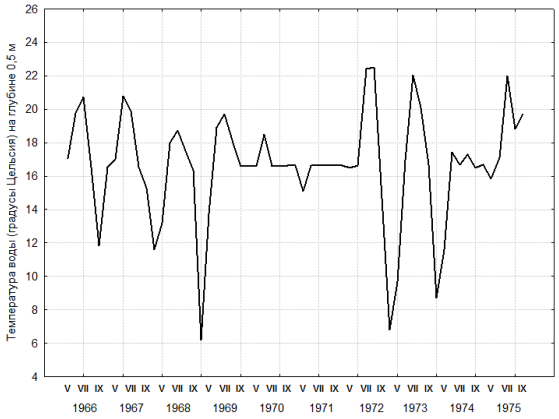

Рисунок 5. График изменения температуры с заполнением пропусков методом линейной регрессии

Заполнение пропусков с помощью расчета процента от знаковой величины

В основе метода лежит расчет процента, какой составляет, в среднем, искомая величина от выбранной знаковой величины (например, максимума или минимума). Этот метод хорош тем, что позволяет учесть как сезонную составляющую, так и тренд. Пример заполнения пропусков таким методом представлен на рисунке 6.

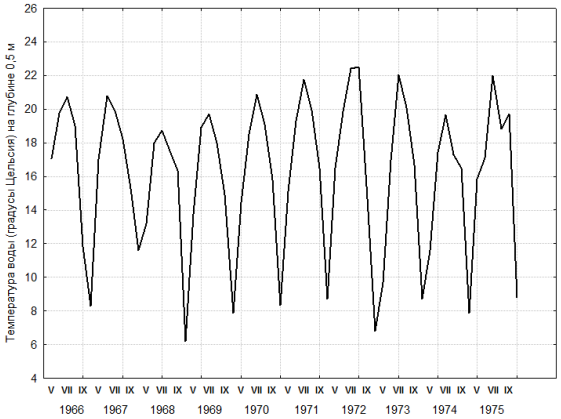

Рисунок 6. График изменения температуры с заполнением пропусков с помощью расчета процента от знаковой величины

Таким образом, для заполнения пропусков во временных рядах с ярко выраженной сезонной составляющей из изученных методов наиболее приемлимо использовать метод заполнения средними значениями по соответствующей фазе колебания, для данных с трендом эффективны методы линейной регрессии, метод интерполяции соседних точек, также могут быть приемлемы методы заполнения средним или медианой из N соседних значений. Для данных, где присудствует и тренд, и сезонные колебания из изученных наиболее подходящим является метод процента от знаковой величины.

ЛИТЕРАТУРА

1. Мальцев К.А., Мухарамова С.С./ Статистический анализ данных в экологии и природопользовании / Казань: Казанский (Приволжский) федеральный университет, 2011г.
2. Круглов В.В., Абраменкова И.В. Методы восстановления пропусков в массивах данных / Программные продукты и системы № 2, 2005 г.
3. Злоба Е., Яцкив И. / Статистические методы восстановления пропущенных данных/ Рига: Институт транспорта и связи, 2002 г.