ГЕНЕТИЧЕСКИЙ АЛГОРИТМ В ЗАДАЧЕ РЕКОНСТРУКЦИИ ГИДРОМЕТЕОРОЛОГИЧЕСКИХ ПОЛЕЙ

Е.Ф. Васечкина, В.Д. Ярин

Морской гидрофизический институт
НАН Украины
г. Севастополь, ул. Капитанская, 2
E-mail: evasvyar@bios.iuf.net

Рассматривается задача восстановления пропусков в массивах данных наблюдений с использованием аппарата эмпирических ортогональных функций. Расчет коэффициентов разложения по этим функциям для массивов со значительным количеством пропусков осуществляется с применением генетического алгоритма. Проведенные расчеты показали высокую эффективность такого подхода в сравнении с традиционными методами, особенно при большом количестве (до 90%) пропусков. Приводятся результаты реконструкции полей приземного давления над северным полушарием по данным за столетний период.

К настоящему времени накоплены массивы гидрометеорологических данных за многие годы. К сожалению, в этих данных имеются многочисленные пропуски (например, участки облачности в спутниковых измерениях поверхностной температуры), в связи с чем возникает задача восстановления полей. Метод разложения полей по ЭОФ, введенный в геофизику Лоренцем в 1956г. [1], имеет уже большую историю применения. Это связано с рядом достоинств этого метода, основными из которых являются большая скорость сходимости рядов по системе ЭОФ в сравнении со всеми другими ортогональными базисами, возможность малопараметрического представления и «естественной» фильтрации полей. Обширную библиографию по использованию ЭОФ анализа в гидрофизике можно найти, например, в работе [2].

Используя аппарат ЭОФ, можно эффективно восстанавливать пропущенные значения при условии достаточности их общего количества. В работе [3] рассматривались вопросы использования метода группового учета аргументов (МГУА) для расчета коэффициентов разложения полей кислорода в Азовском море с целью прогноза динамики поля и восстановления пропущенных данных. Использование МГУА дает значительно меньшую (как минимум в два раза) ошибку прогноза по сравнению с другими методами. Практическое использование этого метода ограничивается требованием достаточности количества данных для обучения модели. В работах [4,5] рассматривалась проблема восстановления профилей гидрохимических характеристик, содержащих большое число пропусков (до 70%). Для восстановления использовалась вся доступная информация о характере распределения этих характеристик по вертикали. По всему имеющемуся массиву данных рассчитывалась ковариационная матрица, которая затем использовалась для расчета набора ЭОФ.

Известно, что произвольная реализация наблюдаемого поля $f(\mathbf{x},t)$ может быть представлена в виде:

$$f(\mathbf{x},t) = f_0(\mathbf{x},t) + f'(\mathbf{x},t)$$
$$f'(\mathbf{x},t) = \sum_k a_k(t)\psi_k(\mathbf{x}) \qquad (1)$$

где ψ_k – собственные функции ковариационной матрицы $P(\mathbf{x},\mathbf{y}) = \sum_t f'(\mathbf{x},t) f'(\mathbf{y},t)$,

$f_0(x,t)$ – математическое ожидание $f(x,t)$.

Коэффициенты разложения $a_k(t)$ для каждой из реализаций рассчитываются путем свертки этой реализации с набором ЭОФ:

$$a_k(t) = \int_s f'(\mathbf{x},t)\psi_k(\mathbf{x})d\mathbf{x} \qquad (2)$$

Если в реализации $f(\mathbf{x},t)$ имеются пропуски, непосредственное применение формулы (2) невозможно. В работе [5] сравнивались различные способы определения коэффициентов разложения по ЭОФ, которые не могут быть рассчитаны стандартным способом, если в реализации имеются пропуски. Было показано, что подбор коэффициентов с использованием генетического алгоритма (ГА) дает лучшие результаты по сравнению с другими методами. А сам способ восстановления профилей путем комбинации аппарата ЭОФ и ГА при большом количестве пропусков существенно точнее сплайновой или линейной интерполяции.

В настоящей работе этот же метод [5] применялся для восстановления пропущенных значений в двумерных реализациях. С его помощью производилась реконструкция массива полей среднемесячного приземного давления в Северном полушарии за 1891 – 1990 гг.[6], содержащего большое количество пропусков.

Алгоритм поиска коэффициентов разложения по системе ЭОФ.

Генетический алгоритм (ГА) относится к группе эволюционных алгоритмов, в основе которых лежит идея поиска решения задачи путем имитации его «биологической» эволюции [7, 8]. Алгоритм оперирует с «популяцией» потенциальных решений исследуемой проблемы. Некоторые из этих решений используются для создания новых решений путем применения к ним специальных операторов. Это операторы селекции, скрещивания (кроссовера) и мутации. Потенциальные решения для оператора скрещивания отбираются на основе оценки их функции качества согласно введенным

критериям, определяемым конкретной задачей. Решения, обладающие лучшим качеством в смысле соответствия введенным критериям, имеют большие шансы «дать потомков» и пройти селекцию для попадания в новое поколение. Новые поколения популяции потенциальных решений рассчитываются последовательно до выполнения некоторого условия, обычно связанного с глубиной найденного экстремума функции качества.

Эффективность алгоритма зависит от многих аспектов, среди которых важное место занимает формализация задачи и способ кодирования ее потенциального решения в виде символьной строки - «генотипа» экземпляра популяции. Решение этого вопроса зависит от специфики задачи, структуры имеющихся для ее решения данных и опыта исследователя. Выбранное математическое описание генотипа определяет в дальнейшем алгоритмические особенности применения эволюционных операторов кроссовера и мутации. Генотип должен однозначно преобразовываться в обычное представление решения задачи, удобное и понятное исследователю. Такое представление в терминах эволюционного моделирования называется «фенотипом» индивидуума популяции. Функция качества индивидуума может быть вычислена только после преобразования его генотипа в фенотип с помощью соответствующей процедуры.

В настоящей работе ГА использовался для поиска набора коэффициентов разложения для каждой реализации, имеющей пропуски, с использованием первых двадцати ЭОФ, совокупность которых определяет 95 % изменчивости поля. Для представления генотипа было применено бинарное кодирование, при котором искомые коэффициенты в двоичном представлении записывались последовательно друг за другом в битовую строку. Таким образом, генотип представлял собой двоичную последовательность, разделенную на равные участки по числу коэффициентов, каждый из которых мог быть преобразован в десятичное число – искомый коэффициент. Первый бит каждого участка определял знак коэффициента.

В начальный момент популяция генерировалась как набор случайных битовых строк заданной длины. Длина генотипа определялась предполагаемыми границами изменчивости коэффициентов, их количеством и точностью вычислений. Общая длина генотипа для поиска коэффициентов при первых N ЭОФ составляла KN двоичных символов, где K – число бит, выделенных для кодирования одного коэффициента. Фенотип экземпляра популяции представлял собой собственно искомое поле давления и рассчитывался по формуле (1).

В алгоритме использовался оператор одноточечного кроссовера, действие которого заключалось в том, что отобранные для скрещивания два экземпляра популяции в случайной точке обменивались своими частями: к началу первого добавлялся конец второго и наоборот. В результате применения оператора кроссовера число экземпляров в популяции увеличивалось вдвое. К каждому новому экземпляру популяции применялся оператор мутации, который с некоторой малой вероятностью мог изменить значение любого бита строки генотипа на противоположное.

Из $2M$ экземпляров («родителей» и «потомков») отбирались M экземпляров, из которых формировалось новое поколение популяции. Отбор производился с вероятностью, пропорциональной функции качества индивидуума. Наилучшее решение из предыдущего этапа сохранялось в новом поколении вне конкуренции. Для предотвращения преждевременного выравнивания популяции по функции качества осуществлялось слежение за ее градиентом. При уменьшении этого градиента до определенного малого значения (вырождение популяции) производилась процедура элиминации, при которой все экземпляры популяции, за исключением нескольких лучших, заменялись новыми, генерируемыми с помощью датчика случайных чисел.

Определяющие параметры ГА - размер популяции, количество поколений, вероятность мутации, длина строки генотипа - варьировались в довольно широких пределах. Как показали расчеты, оптимальные значения перечисленных параметров находятся в следующих пределах: размер популяции 200–250, число генераций 100–150, количество оцениваемых коэффициентов (число ЭОФ, используемых для аппроксимации) 15–20, точность оценки коэффициентов 10^{-4}–10^{-5}, длина генотипа соответственно 500–560 символов, вероятность мутации 0.01–0.03.

Основной проблемой при применении ГА является адекватный выбор функции качества для оценки потенциальных решений. В данном случае наряду с требованием минимума среднеквадратической ошибки аппроксимации поля в точках сетки, где имелись измерения, к функции качества был добавлен член, зависящий от второй производной восстанавливаемой реализации. Таким образом, предпочтение отдавалось более гладким восстанавливаемым распределениям приземного давления.

Численные эксперименты.

Данные по полю приземного давления в Северном полушарии представляют собой аномалии среднемесячного давления на сетке с шагом 5^0 по широте и 10^0 по долготе за столет-

ний период с 1891г. Данные до 1964 г. содержали значительное число пропусков (от 20 до 70%). Три широтных круга, соответствующие приэкваториальным областям, были исключены из рассмотрения ввиду отсутствия данных. Таким образом, в экспериментах использовался массив размерностью 36 x 16 x 1124.

Были выполнены два численных эксперимента: в первом ковариационная матрица рассчитывалась с использованием только полных реализаций, не имеющих пропусков; во втором в эти же реализации были искусственно внесены 35% пропусков, распределенных случайным образом. Обе ковариационные матрицы затем использовались для расчета ЭОФ и в алгоритме восстановления полей.

I. Из 1124 реализаций только 177 не имели пропусков и могли быть сразу использованы для расчета ковариационной матрицы. Для проверки точности восстановления полей 17 реализаций из 177 были выделены в качестве контрольной последовательности, а по остальным 160 были рассчитаны первый и второй моменты распределения и система ЭОФ.

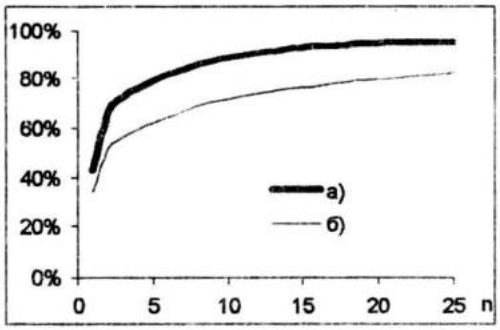


Рис.1. Интегральные кривые роста энергии с ростом числа ЭОФ в случае, когда ковариационная матрица рассчитывалась: а) по полным данным; б) – по реализациям с 35% пропусков.

На рис.1 жирной линией показана интегральная кривая роста энергии при учете последовательных мод (распределение собственных чисел по модам ЭОФ см. на рис.7.). Легко видеть, что первые пять мод несут в себе почти 80% всей изменчивости поля: первая - 43%, вторая – 24%, третья – 5%, далее кривая распределения выравнивается. Вопрос о достаточном количестве используемых функций решается в зависимости от специфики каждой конкретной задачи. Для определенности мы ограничились первыми 20 ЭОФ, описывающими 95% изменчивости поля.

Расчеты проводились следующим образом. По 160 реализациям поля давления были рассчитаны ЭОФ, а оставшиеся 17 использовались в качестве контрольной выборки, на которой оценивалась точность реконструкции полей. Расположение точек на сетке, содержащих пропущенные значения, задавалось двумя различными способами: в первом случае позиции этих точек задавались с помощью датчика случайных чисел; во втором имитировалось наличие данных о поле приземного давления в 1891-1900гг. (годы, наименее обеспеченные данными). Расчетная сетка с размещением на ней точек с данными для двух этих случаев, представлена на рис.2.

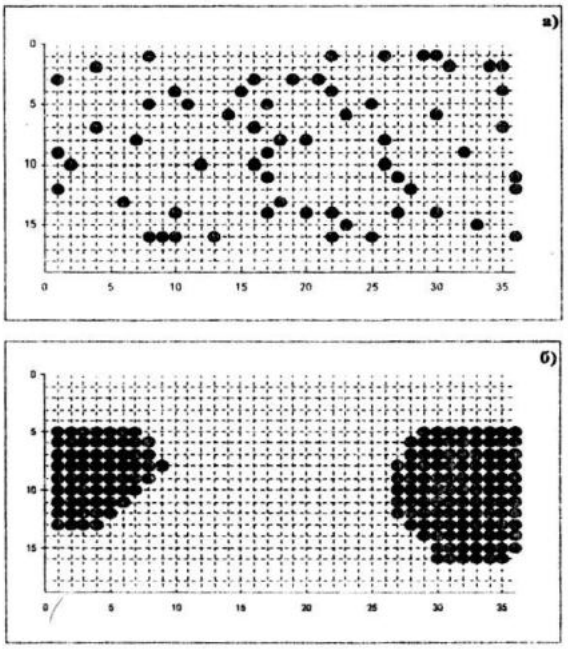
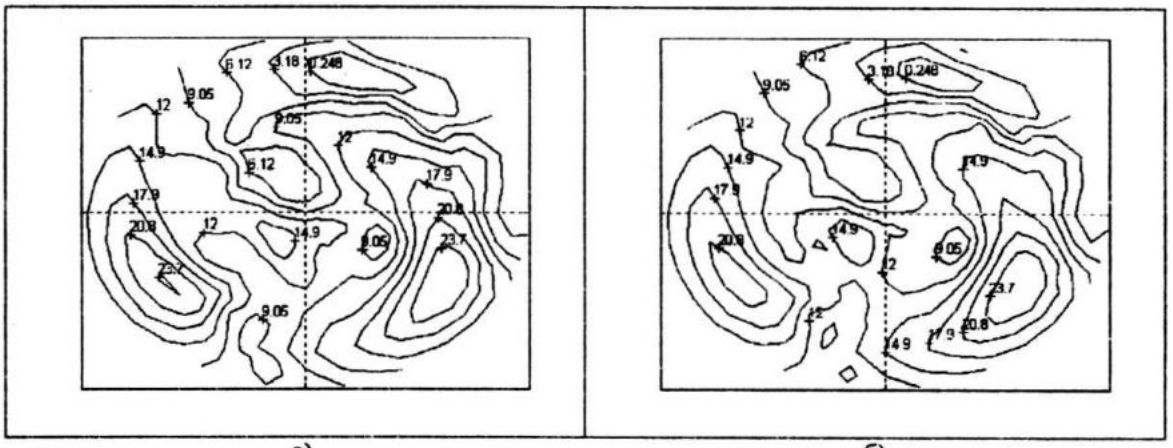


Рис.2. Расчетная сетка с обозначением точек, в которых имелась информация о приземном давлении: а) 60 точек случайным образом распределены по сетке; б) информация о поле давления локализована в одном районе (168 точек).

Результаты реконструкции поля с минимальной и максимальной ошибками восстановления представлены на рис. 3 - 5. При поиске коэффициентов разложения использовались точные значения приземного давления, а результаты реконструкции сравнивались со сглаженными реализациями. Ошибка рассчитывалась как среднеквадратическая разность между восстановленным и истинным, сглаженным по 20 функциям, полями во всех точках сетки. Средняя по всей контрольной последовательности ошибка восстановления полей при 90% пропусков, распределенных случайным образом, составила 1,9мбар (35,3%). Относительная ошибка рассчитывалась как отношение среднеквадратической ошибки аппроксимации к стандартному отклонению сглаженных реализаций выборки.

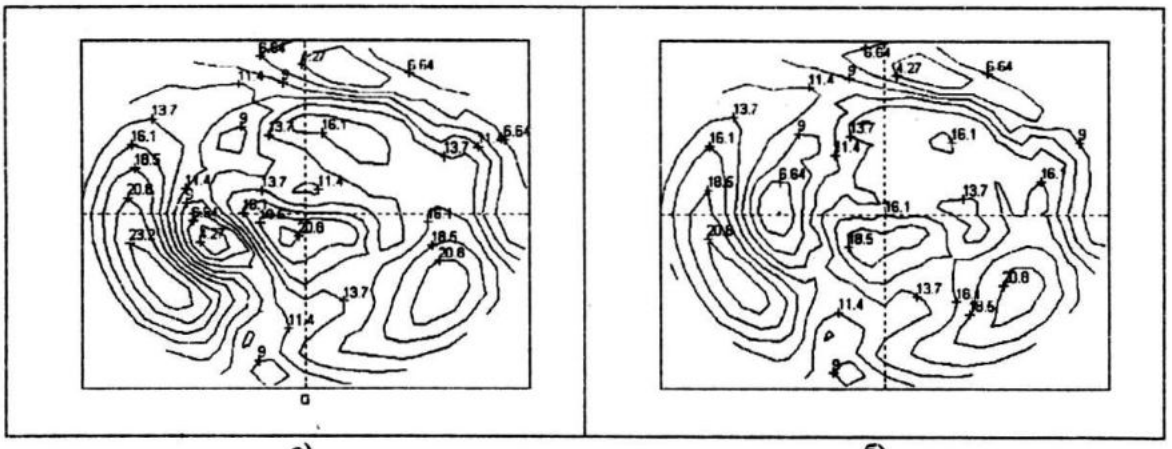
При наличии информации, сосредоточенной в одной локальной области, результаты получились хуже, средняя ошибка составила 56,7% (3,1мбар). Результаты реконструкции иллюстрируют рис. 4 и 5. Несмотря на большую среднеквадратическую ошибку поля на

рис. 5в, локализация областей пониженного и повышенного давления показана правильно, Гистограммы ошибок реконструкции полей для двух вариантов расположения пропусков представлены на рис. 6. Как видим, и в том и в другом случаях медиана распределения смещена в сторону меньших величин ошибки.



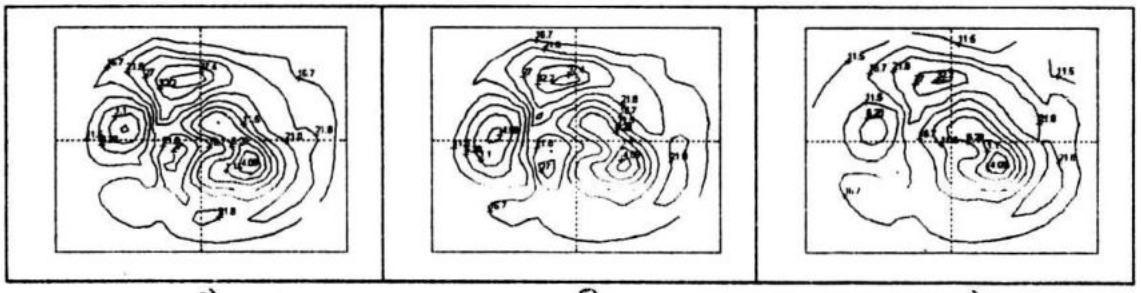
а) б)

Рис.3. Результат расчета с минимальной дисперсией, полученный при 90% пропусков, случайным образом распределенных на сетке (рис. 2а): а) истинное поле; б) восстановленное.



а) б)

Рис.4. Результат расчета с минимальной дисперсией, полученный при наличии информации лишь в ограниченной области (рис. 2б): а) истинное поле; б) восстановленное.



а) б) в)

Рис.5. Результат расчета по восстановлению полей давления с максимальной дисперсией ошибки: а) истинное поле; б) восстановленное поле при 90% пропусков, случайным образом распределенных на сетке (рис. 2а); в) восстановленное поле при наличии информации лишь в ограниченной области (рис. 2б).

II. Для имитации реальной ситуации, ковариационная матрица рассчитывалась также по реализациям, в каждую из которых искусственно вносились пропуски (до 35%). Поскольку в таком случае каждый ее элемент был неодинаково обеспечен данными, это отразилось на виде ЭОФ. На рис.7 тонкой линией показано распределение энергии по модам ЭОФ, рассчитанных по неполным данным, в сравнении с ЭОФ «без пропусков». Рис.1 (также тонкая ли-

ния) показывает поведение интегральной кривой роста энергии. Видно, что первые моды несут в себе меньше энергии, первые 20 ЭОФ всего 80% вместо 95%, как в предыдущем расчете. Различается и внешний вид функций, причем это различие растет с номером моды, см. рис. 8.

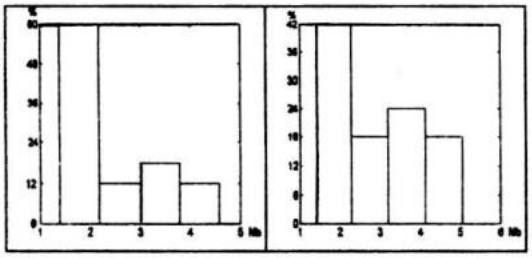


Рис.6. Гистограммы среднеквадратической ошибки реконструкции полей приземного давления (контрольная выборка): а) случай размещения пропусков с помощью датчика случайных чисел; б) случай локализации информации о поле в одной области.

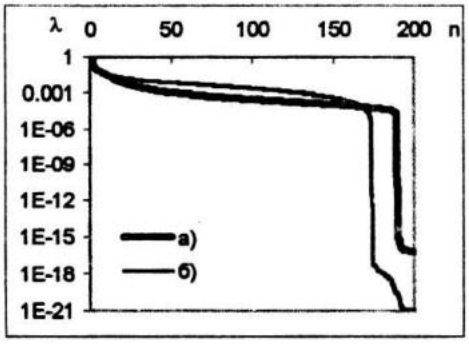


Рис.7. Распределение дисперсии изменчивости поля давления по модам ЭОФ в случае, когда ковариационная матрица рассчитывалась: а) по полным данным; б) – по реализациям с 35% пропусков.

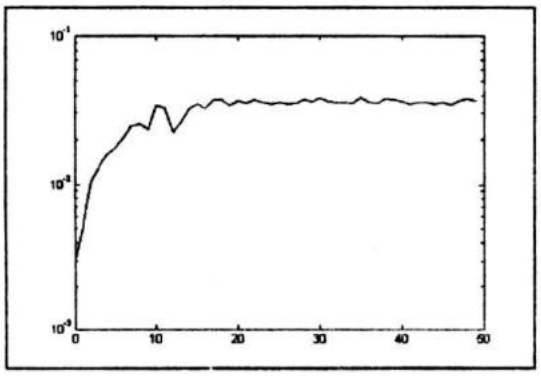


Рис.8. Распределение среднеквадратической ошибки расчета ЭОФ по неполным данным в сравнении с ЭОФ, рассчитанным по массиву реализаций без пропусков.

Из рис.8 становится очевидным, что в условиях неполноты данных для расчета ковариационной матрицы нужно использовать меньшее число функций для восстановления полей. Проведенные расчеты действительно показали, что в этом случае с уменьшением числа используемых ЭОФ среднеквадратическая ошибка восстановления полей уменьшается, и оптимальным оказывается использование только 10 функций. Ошибка восстановления полей по 10 ЭОФ при 90% пропусков составила 2,5мбар (46%), что примерно соответствует ошибке восстановления с использованием ковариационной матрицы, рассчитанной по реализациям без пропусков при том же числе функций - 2,1мбар (39%).

Заключение. Выполненные эксперименты позволили сделать вывод, что предложенный метод дает возможность реконструировать поля гидрометеорологических элементов с удовлетворительной точностью даже при наличии данных всего в 10% узлов сетки. Этот метод был нами использован для реконструкции среднемесячных полей приземного давления с 1891г. по 1963г., имеющих значительное количество пропусков.

ЛИТЕРАТУРА

1. Lorenz E. Empirical orthogonal functions and statistical weather prediction // Tech.Rep.1, Statistical Forecasting Project, Department of Meteorology, Massachusetts Institute of Technology. - Cambridge, Massachusetts. - 49pp.
2. Everson R., Cornillon P., Sirovich L., and Webber A. An empirical eigenfunction analysis of sea surface temperatures in the Western North Atlantic // J. of Phys.Oceanogr. – 1997. - v.27. - № 3 – P.468-479.
3. Васечкина Е.Ф., Тимченко И.Е., Ярин В.Д. Восстановление структуры гидрофизических полей по неполной информации // МГФЖ. – 1997. - № 2. - С.37-44.
4. Васечкина Е.Ф., Ярин В.Д. Применение генетических алгоритмов в математическом моделировании экосистем // Материалы Международного научно-технического семинара «Системы контроля окружающей среды - 2000». – Севастополь. – 2000. - С.346-355.
5. Васечкина Е.Ф., Ярин В.Д. Использование генетического алгоритма для восстановления пропущенных данных // МГФЖ – в печати.
6. Банк данных МГИ НАН Украины. URL: http://www.mhi.iuf.net/ DEPTS.
7. Beasley D., Bull D.R., Martin R.R. An overview of Genetic Algorithms: Part I, Fundamentals. // University Computing – 1993. – v.15. - № 2. – P.58-69.
8. Beasley D., Bull D.R., Martin R.R., An overview of Genetic Algorithms: Part II, Research Topics. // University Computing – 1993. – v. 15. - №4. – P.170-181.