Лекции БШФФ-2006. С. 39–46

УДК 551.510.535+519.25

СОВРЕМЕННЫЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ В ГЕОФИЗИКЕ

А.Н. Ляхов

Введение

Современная геофизика немыслима без анализа экспериментальных данных. При этом состояние среды является суперпозицией очень большого количества нелинейных взаимодействий между разными процессами или, еще можно сказать, разными степенями свободы, или *модами*. Последний термин будет сопровождать нас на протяжении всей лекции. Построение физической картины происходящих процессов можно начинать и из первопринципов, полагаясь на мощь суперкомпьютеров. Но в этом случае объем выданной расчетной информации становится сопоставим с реальной средой и вопрос интерпретации полученных результатов остается на повестке дня. Методы, которые будут рассмотрены в данной лекции, применимы для анализа расчетной информации точно так же, как и для анализа наблюдательных данных.

Первая задача, которая стоит перед исследователем, анализирующим геофизические поля данных, найти способ уменьшить размерность системы и выявить *структуры*, наиболее полно объясняющие наблюдаемые вариации параметров.

Следующий этап работы – взаимный анализ поведения выявленных структур в разных полях данных или выявление и описание отклика структур на вариации внешних, «управляющих» (или подозрительных с этой точки зрения) параметров.

В дальнейшем изложении мы исходим из посылки, что собственно измерения выполнены с известной точностью и мы, во всяком случае, имеем оценку среднего значения наблюдаемой величины и ее дисперсию. В процессе изложения детально будут рассмотрены подходы, построенные на использовании только средних значений величин. Модификация методов для интервально заданных данных и для включения известных дисперсий описана в литературе, приведенной в конце.

Наша лекция посвящена анализу пространственно-временных данных методом естественных (эмпирических) ортогональных функций (ЕОФ) и его разновидностям. Этот метод стал за последние десятилетия одним из основных рабочих инструментов в метеорологии, физике атмосферы и океана. Отдельно рассмотрим вариант метода для анализа (в том числе взаимного) временных рядов (сингулярный спектральный анализ и его многоканальный вариант). Демонстрация практического применения метода для задач физики ионосферы будет представлена анализом полей данных полного электронного содержания и данных мировой сети магнитометрических станций.

Метод естественных ортогональных функций: пространственные структуры

Метод используется для анализа карт полей данных скалярных величин, меняющихся во времени. Метод находит пространственные структуры вариации сигнала, их динамику во времени и «измеряет» относительную важность каждой структуры. Возможен совместный анализ вариации двух и более полей, хотя этот подход не получил широкого распространения из-за сложностей с интерпретацией полученных результатов.

Основная идея, лежащая в основе аппарата ЕОФ, следующая. Предположим, что мы можем изобразить каждую из карт данных в виде вектора f_m в N-мерном пространстве для каждого момента времени n. Все вектора исходят из начала координат в M-пространстве. Если исходные данные скоррелированы, то вектора будут сгущаться (кластеризоваться) вдоль каких-то выделенных направлений. Проблема, решаемая с помощью метода ЕОФ, заключается в поиске такого ортогонального базиса $\{e_1, e_2, \ldots, e_N\}$ (системы координат) в N-мерном пространстве, чтоб вектор e_1 был направлен к самому большому кластеру, вектор e_2 к следующему по величине и т. д. Так что сумма квадратов проекций всех векторов f на направления $\{e_1, e_2, \ldots, e_N\}$ убывает строго последовательно. Из-за ортогональности векторов $\{e_1, e_2, \ldots, e_N\}$ найденные структуры называются ортогональными функциями. А так как новый базис строится по самим данным, а не выбирается априорно, эти функции называются естественными (в английской литературе – эмпирическими).

Очень важно! Метод расщепляет исходные поля данных на «моды данных». Эти моды – не всегда (!) совпадают с физическими. Один и тот же физический процесс может давать вклад в разные моды, и одна и та же мода может быть результатом действия более чем одного физического процесса.

Результат применения метода позволяет упростить наблюдаемую картину, сделать ее более легкой для интерпретации. Человеку затруднительно и часто просто невозможно проанализировать данные высокой размерности (4D – широта, долгота, параметр и время!). Даже визуализация таких данных – нетривиальная задача.

Небольшое отступление о терминологии.

В литературе используются названия метод главных компонент (principal component analysis) и метод естественных ортогональных функций (empirical orthogonal functions). Математический аппарат этих методов совпадает, а разнобой в терминологии вызван историческими причинами. Мы будем использовать термин естественные ортогональные функции (далее по тексту ЕОФ).

Кроме собственно ЕОФ, существуют его разновидности, из которых мы рассмотрим следующие: расширенные ЕОФ (extended EOF) – включают информацию об авто- и кросскорреляционных свойствах, частотные ЕОФ (frequency domain EOF) и ком-

плексные ЕОФ (Hilbert EOF) – все для анализа пространственно-временных структур; метод сингулярного спектрального анализа (SSA) или временные ЕОФ для анализа временных последовательностей.

Организация данных и их предварительная обработка

Пусть имеется некоторая сеть наблюдений, характеризующаяся набором координат $\{x_i, y_i\}$, $i = 1...N$. То есть мы имеем N пунктов наблюдений. Это могут быть магнитометрические станции, сеть ионозондов, узлы сетки, в которых определено полное электронное содержание, или какие-либо метеорологические параметры – в принципе, это может быть любая измеренная величина. Пока ограничимся только информацией об абсолютной величине параметра. Роль ошибки измерения обсудим ниже. Каждая станция ведет наблюдения в течение некоторого времени, так что на каждой станции сформирован вектор наблюдений для отсчетов времени t_i, $i = 1...M$.

Необходимым условием применения рассматриваемых методов является синхронизация измерений. Провалы в данных должны быть заполнены каким-либо способом, либо должны использоваться только данные на синхронные моменты времени. Существующие эффективные алгоритмы заполнения пропусков основаны, в свою очередь, как раз на рассматриваемом методе, так что слепое их использование приведет к тавтологии в результатах.

Имеющиеся данные необходимо организовать в виде матрицы следующего вида:

$$F = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{bmatrix}. \qquad (1.1)$$

Первая строка матрицы F – это «мгновенный снимок» данных по всем станциям в момент времени t_1, а первая колонка – это временной ряд наблюдений по станции № 1 и т.д. Алгоритм упорядочивания станций в строке значения не имеет, важно только зафиксировать свой выбор и использовать его далее при обратном восстановлении карт полей.

При обработке данных, заданных на однородных географических сетках (с фиксированным шагом по широте и долготе), наблюдается сгущение точек в высоких широтах. Этот эффект может влиять на структуры ЕОФ. Чтобы избежать ложных выводов, необходимо скорректировать данные. Аккуратный подход заключается в разбиении поверхности Земли на сегменты равной площади, с усреднением исходных данных по этим сегментам. В этом случае элементы $x_{i,j}$ в (1.1) будут соответствовать номеру сегмента. Более простой, но достаточно корректный метод заключается в умножении всех данных на $\cos\vartheta_i$, где ϑ_i – широта станции наблюдений.

Алгоритм расчета классических ЕОФ

Второй шаг вычислений – рассчитываем среднее по столбцам $\overline{x} = \{\overline{x}_1, \ldots, \overline{x}_N\}$ и вычитаем средние значения из столбцов матрицы F.

Третий шаг вычислений – расчет сингулярного разложения матрицы F. Любую матрицу A размером $M \times N$ можно представить в виде $A = U \cdot \Sigma \cdot V^T$, с матрицей U размера $M \times M$, диагональной матрицей Σ размера $M \times N$ и матрицей V размера $N \times N$. Матрица Σ содержит неотрицательные элементы σ только на главной диагонали, называемые сингулярными значениями матрицы A. Элементы на главной диагонали убывают по величине. Относительный вклад в полную вариацию от j-го сингулярного значения определяется как

$$\sigma_j^2 \Big/ \sum_{k=1}^{N} \sigma_k^2 .$$

Максимальный номер ненулевого сингулярного значения определяет максимально возможное число получаемых ЕОФ. Столбцы матрицы U ортогональны и называются левыми сингулярными векторами матрицы A, они и являются естественными ортогональными функциями (пространственными модами), соответствующими данным сингулярным значениям. Строки матрицы V^T также ортогональны и называются правыми сингулярными векторами матрицы A. Они пропорциональны главным компонентам (временным) матрицы A:

$$PC = \Sigma \cdot V^T, \qquad (1.2)$$

$$PC_k(t) = \sigma_k V^{Tk}(t). \qquad (1.3)$$

Исходное поле наблюдений F может быть восстановлено следующим образом:

$$F_m(t) = \sum_{k=1}^{L} U_m^k \cdot \sigma_k \cdot V^{Tk}(t), \qquad (1.4)$$

где L – число удерживаемых сингулярных значений, выбранное исходя из объяснимого уровня вариации. При удержании всех ненулевых сингулярных значений мы восстановим исходную картину.

Чтобы проанализировать пространственное распределение первой моды ЕОФ, необходимо взять столбец матрицы U, соответствующий первому сингулярному значению. Его «раскладка» по станциям наблюдений будет совпадать с использованной при построении матрицы F. Выполнив необходимые операции, мы получим пространственную карту первой моды данных.

Длина вектора главной компоненты совпадает с длиной анализируемой записи. Эта временная динамика может изучаться классическими спектральными методами или пропускаться через кросскорреляционный анализ.

Некоторые мифы ЕОФ

Миф 1. ЕОФ – это разновидность факторного анализа (ФА).

- ЕОФ объясняет наблюдения, но не базируется на модели данных.
- ЕОФ концентрируется на объяснении вариации, ФА – на корреляциях.
- Выбор числа факторов важнее в ФА.
- По умолчанию выполняется вращение факторов, что не верно для ЕОФ.

Миф 2. Данные должны иметь нормальное распределение.

Иногда – да, но в целом – нет!

ЕОФ – это описательный метод. Лишь в редких случаях, когда требуется формальный вывод, нам необходимы допущения о законах распределения. Исключение – категорийные данные (тип облачности, знак отклонения и т.д.), здесь вопрос остается открытым.

Миф 3а. ЕОФ дают физические моды.

Миф 3б. ЕОФ не может дать физические моды.

ЕОФ выдает линейные моды максимальной вариации. Являются ли эти моды физическими или нет, зависит от того, что мы вкладываем в понятие «физические». В любом случае моды, объясняющие максимум вариации, должны быть интересны для анализа. В целом, конечно, можно постулировать, что а) несколько физических процессов могут давать вклад в одну и ту же моду и б) один и тот же физический процесс может давать вклад в несколько разных мод.

Миф 4. Число удерживаемых ЕОФ очень важно.

Простого и ясного правила не существует.

Миф 5. Высшие моды содержат шум.

Не исключено, хотя моды малой вариации могут быть полезны.

- Они минимизируют вариацию и могут детектировать изначально неподозрительные квазипостоянные связи между переменными.
- Они могут быть полезны для детектирования выбросов в данных.
- Если ЕОФ используется для предварительной обработки данных, после чего используется другой метод анализа (например спектральный), то высшие ЕОФ могут быть более полезны, чем первые.

Еще об интерпретации результатов

Первое на что мы смотрим – это доля объясненной вариации. Неявно полагается, что большая вариация, связанная с доминирующей модой, имеет физическое объяснение. Ключ к пониманию смысла найденной моды часто лежит в анализе временной динамики главной компоненты $PC_k(t)$. Сами пространственные структуры ЕОФ конструируются так, чтобы оптимально представить вариацию поля данных, а не физические связи или максимальные корреляции. Поэтому даже моды, объясняющие значимую вариацию, могут не иметь ничего общего с физической картиной явлений.

Дополнительная сложность вызвана ортогональностью ЕОФ и независимостью временных главных компонент. Часто бывает так, что первая мода легко увязывается с физическим процессом. Но вторая мода обязана быть ортогональна к первой! А значит, вторая мода должна описывать процесс, *действующий независимо* от первого. Природные процессы, тем не менее, не всегда независимы и чаще взаимосвязаны.

Традиционный метод ЕОФ позволяет выявлять только стоячие колебания. В некоторых случаях признаком того, что мы имеем дело с распространяющимся сигналом, может служить ситуация, когда две главных компоненты $PC_k(t)$ и $PC_{k+1}(t)$, соответствующие сингулярным значениям σ_k, σ_{k+1}, меняются когерентно со сдвигом по фазе на $\pi/2$. В этом случае сингулярные значения близки по величине и данная пара ЕОФ и главных компонент может представлять сигнал, распространяющийся в пространстве. В этом случае необходимо использовать другие разновидности метода – расширенные или комплексные ЕОФ.

Временные структуры: сингулярный спектральный анализ (ССА)

Данный вид обработки данных является вариацией метода ЕОФ в приложении к одиночному временному ряду. В методе ЕОФ мы имели сеть станций, измерения на которых были упорядочены во времени в виде матрицы F, и мы выявляли доминирующие пространственные структуры. В методе ССА поле F содержит данные в одной точке, но на разных временных лагах. Собственные вектора соответствующей ковариационной матрицы дадут доминирующие временные структуры.

Начинаем с временного ряда наблюдений $Y(t) = \{Y_0, Y_1, \ldots, Y_N\}$. Сконструируем матрицу F из сдвинутых на временной лаг $l \cdot \Delta$ копий исходного ряда, где $l = 0, \ldots, L$, причем начинаем с $l = 0$. Выбор шага Δ определяется исследователем. Для $\Delta = \Delta t$ матрица F будет выглядеть так:

$$F = \begin{bmatrix} F(1) & F(2) & \ldots & F(N-L) \\ F(2) & F(3) & \ldots & F(N-L+1) \\ F(3) & F(4) & \ldots & F(N-L+2) \\ \vdots & \vdots & \ldots & \ldots\ldots \\ F(L+1) & F(L+2) & \ldots & F(N) \end{bmatrix}. \quad (2.1)$$

Матрица F имеет размер $L + 1 \times N - L$. Далее рассчитываем матрицу ковариации

$$R_{FF} = F * F^T. \quad (2.2)$$

Эта матрица имеет размер $L + 1 \times L + 1$, и ее элементы равны ковариации $\langle Y(t+l)Y(t+l)\rangle$ между копиями временного ряда на всех возможных комбинациях лагов $l = 0, \ldots, L$. Далее решается задача на собственные значения $R_{FF} * E = E * \Lambda$, где Λ – квадратная матрица размера $L + 1 \times L + 1$ с ненулевыми элементами на главной диагонали, а матрица E содержит собственные вектора. В отличие от пространственных ЕОФ, собственные вектора в матрице E теперь содержат *временные структуры*. В западной литературе их называют временными ЕОФ. Каждый собственный вектор E_l^k – это сдвинутая на лаг l временная последовательность длиной $L + 1$, а k – номер моды.

Главные компоненты $TPC_k(t)$ рассчитываются по

$$TPC_k(t) = E^T * F, \quad (2.3)$$

$$TPC_k(t) = \sum_{l=0}^{L} E_l^k F(t+l). \quad (2.4)$$

Эти функции являются отфильтрованными версиями исходного сигнала. При этом сумма спектров $TPC_k(t)$ в точности совпадает со спектром исходной последовательности.

Реконструкция сигнала по выбранной моде осуществляется так:

$$RC_l^k(t) = E_l^k TPC_k(t). \qquad (2.5)$$

В отличие от классического спектрального анализа, в котором функции разложения заданы изначально как набор синусов и косинусов, в ССА эти функции определяются из самих данных, формируют ортогональный базис разложения, оптимальный в статистическом смысле. В ССА любая колебательная динамика в исходном временном ряде детектируется парой близких собственных значений в Λ и колебание можно изолировать: $Z_k(t) = TPC_k(t) + iTPC_{k+1}(t) = = \rho(t)e^{i\psi(t)}$. Благодаря этому свойству, ССА-метод наилучшим образом подходит для изоляции ангармонических колебаний с меняющейся амплитудой в зашумленных данных. Также метод позволяет выявлять даже слабые, но значимые тренды в данных, в том числе нелинейные!

Помимо собственно ССА-метода разработаны его версии MonteCarlo-SSA, позволяющий определять статистическую значимость результатов, и MultiChannel-SSA, совпадающий по смыслу с аппаратом расширенных ЕОФ, которые будут рассмотрены ниже. В многоканальном варианте ищутся ковариирующие структуры в разных временных рядах.

Пространственно-временные структуры

Классический метод ЕОФ, описанный в первом параграфе, позволяет выявлять только стоячие колебания в данных. Он не применим в том случае, когда мы имеем дело с бегущей волной или движущимися структурами. Это сильное ограничение, так как многие физические процессы проявляются в результате взаимодействия бегущих волн разных длин и частот.

Расширенные (Extended EOF) ЕОФ рассчитываются по классической схеме, но матрица *F* формируется более сложным образом. В начале по формуле (1.1) рассчитывается сама матрица *F*. Далее к ней снизу стыкуется ее копия, сдвинутая на лаг *l*, далее снизу добавляется еще одна копия матрицы, сдвинутая на лаг 2*l*, и т.д. Получается очень большая матрица размером $(L+1)*M \times N-L$. Дальнейший расчет аналогичен классическому. После вычислений сингулярного разложения карты ЕОФ содержат не одиночную структуру, но последовательность $L + 1$ сдвинутых карт. Анимация этих карт несет информацию о возможной движущейся структуре, связанной с данной модой.

Частотные (frequency domain EOF) ЕОФ являются одними из наиболее употребительных для анализа движущихся структур. К сожалению, в ситуации, когда энергия моды распределена по спектру, что, в частности, возможно при флуктуациях на нерегулярных интервалах, требуется большое количество карт данных для хорошего описания единичного события. Алгоритм расчета строится по следующей схеме: 1) вычисляем дискретное фурье-преобразование временного ряда в каждой точке наблюдений $F_m(t) \to Y_m(f)$; 2) формируем матрицу *Y*, в колонках которой находятся спектральные коэффициенты для всех станций наблюдений, для частот от 0 до $f_s/2$ (f_s – частота дискретизации), так что теперь одна строка – это энергия на данной частоте по всем станциям. Выполняем сингулярное разложение матрицы $Y \to U \cdot \Sigma \cdot V^T$. Сингулярные вектора U_m^k дают комплексные ЕОФ, а вектора V_m^{kT} – комплексные аналоги главных компонент, но в частотной области – комбинации чисто гармонических компонент, описывающих относительно плавную динамику *k*-й ЕОФ. Сохранение информации о фазе дает возможность получать как стоячие, так и бегущие возмущения.

Комплексные (Гильбертовы) ЕОФ – наиболее современные, избавлены от ограничений частотных ЕОФ и применимы для анализа векторных полей. Если мы анализируем скалярное поле, то на первом шаге данные искусственно комплексифицируются с использованием собственно данных и их преобразования Гильберта. Для векторных полей данные образуются из широтной и долготной компонент: $f = X + iY$. Далее, аналогично вышеизложенному, выполняем разложение комплексной матрицы $\tilde{F}$ и определяем комплексные ЕОФ. Метод позволяет эффективно выявлять перемещающиеся структуры, особенно в случае, если вариация распределена по спектру, что часто наблюдается для природных данных.

Пусть в некоторой точке координат мы имеем временной ряд наблюдений $\varphi_m(t)$. Комплексная величина формируется из этих данных и их преобразования Гильберта:

$$\Phi_m(t) = \varphi_m(t) + i\hat{\varphi}_m(t).$$

Если фурье-представление $\varphi_m(t)$ определить как

$$\varphi_m(t) = \sum_\omega a_m(\omega)\cos\omega t + b_m(\omega)\sin\omega t,$$

то преобразование Гильберта $\hat{\varphi}_m(t)$ поля $\varphi_m(t)$ будет

$$\hat{\varphi}_m(t) = \sum_\omega b_m(\omega)\sin\omega t - a_m(\omega)\cos\omega t. \qquad (1.5)$$

Преобразование Гильберта является операцией фильтрации $\varphi_m(t)$, при которой амплитуда каждой спектральной компоненты остается неизменной, а фаза смещается на $\pi/2$. Комплексифицированные наблюдения $\Phi_m(t)$ упорядочиваются в виде матрицы *F*, как показано выше, после приведения к стандартному виду (удаление среднего значения). После этого выполняем сингулярное разложение и находим ЕОФ и главные компоненты. Обе этих величины будут комплексными. Естественная ортогональная функция может быть представлена через пространственную амплитуду и пространственную фазу:

$$E_m^k = B_m^k \exp(i\Theta_m^k).$$

Аналогично главная компонента представима через временную амплитуду и временную фазу: $A^k(t) = C^k(t)\exp(i\Psi^k(t))$. Четыре величины (простран-

ственная амплитуда, пространственная фаза, временная амплитуда и временная фаза) дают общее представление любых динамических структур в исходном поле данных *F*. Так как они определены в пространстве и во времени, интерпретация результатов не страдает при наличии циклостационарных сигналов. Если в данных присутствуют периодичности, они будут обнаружены этим методом.

Пространственная амплитуда B_m^k показывает распределение в пространстве вариации, связанной с данной собственной модой, а пространственная фаза Θ_m^k показывает относительную фазу флуктуаций для разных точек на карте, на множестве которых определена исходная матрица *F*. Временная амплитуда $C^k(t)$ интерпретируется как главная компонента и описывает временной ход амплитуды данной моды. Временная фаза $\Psi^k(t)$ описывает изменение фазы, связанное с периодичностями в *F*. Реконструкция поля наблюдений осуществляется обычным образом:

$$F_m(t) = \mathrm{Re}\left(\sum_{k=1}^{L} E_m^{k*} \cdot A^k(t) \right).$$

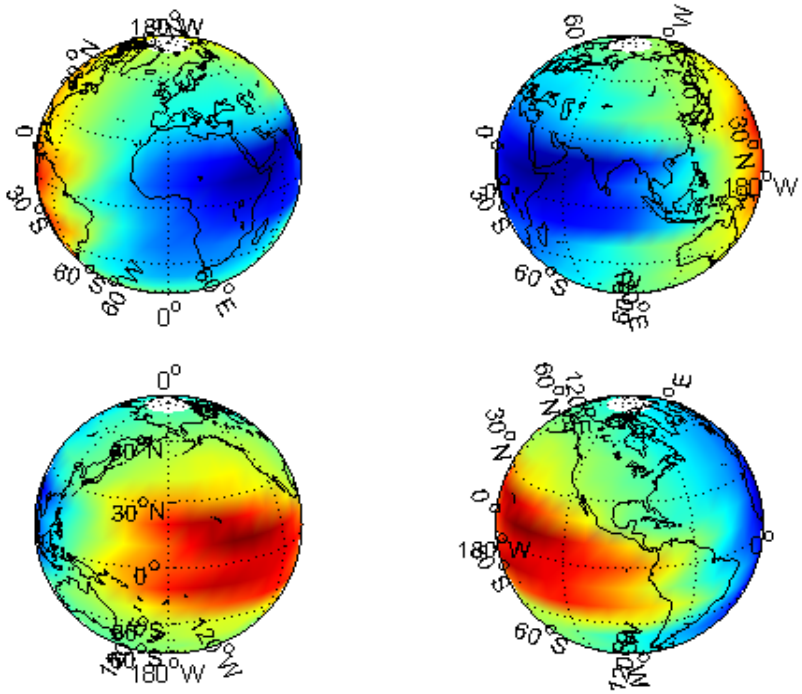

Рис. 1. Первая ЕОФ для поля ТЕС.

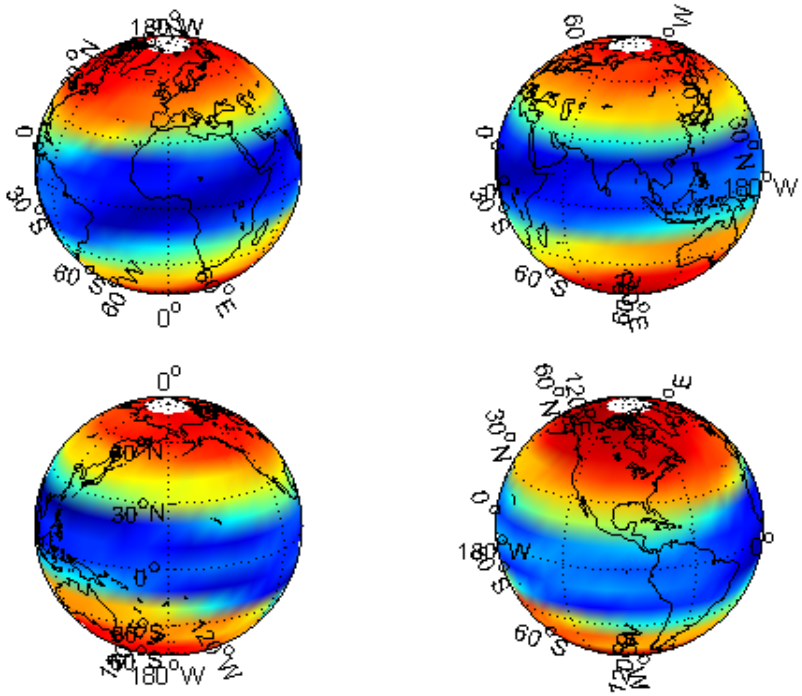

Рис. 2. Вторая ЕОФ для поля ТЕС.

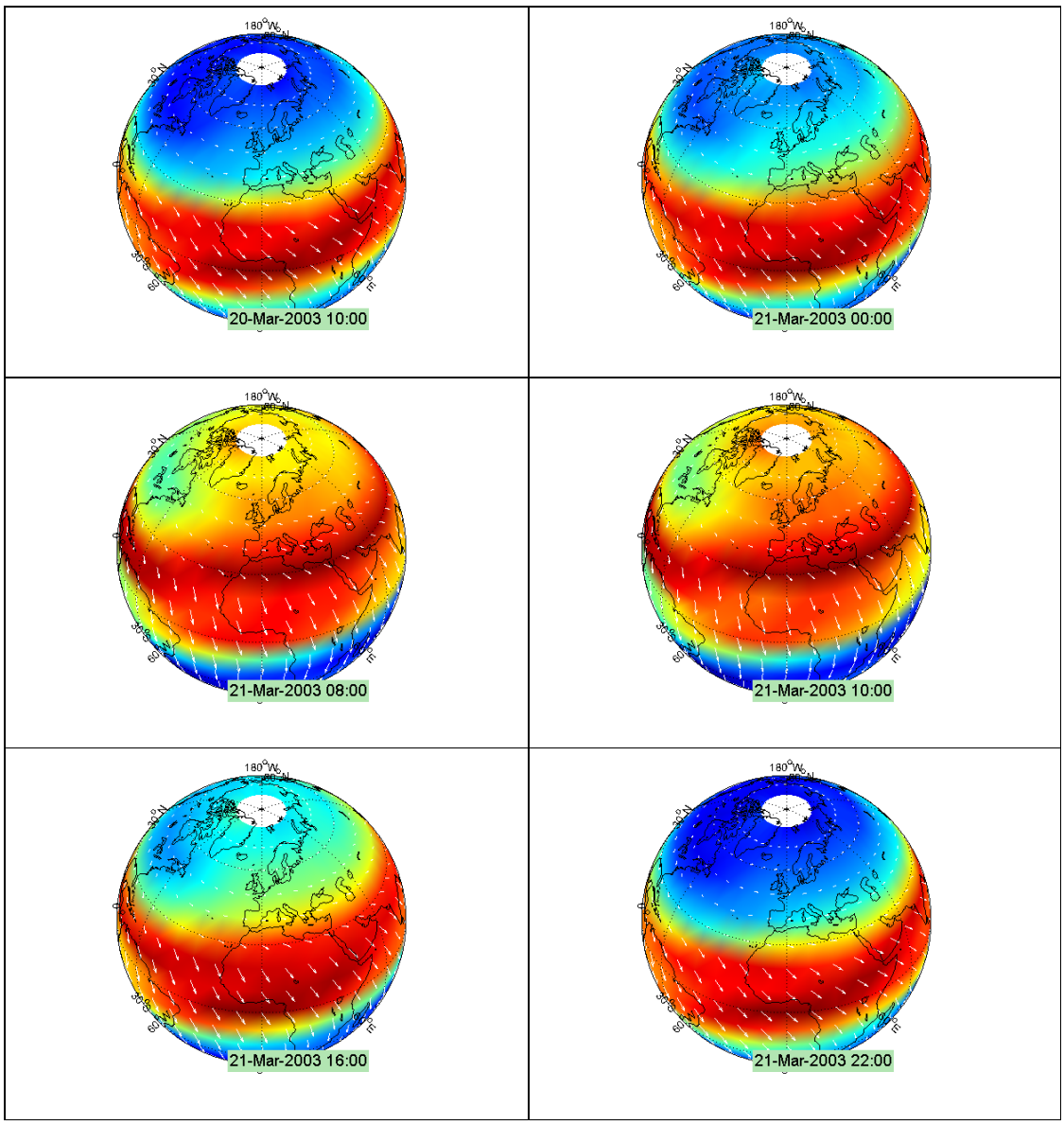

Рис. 3. Отклик второй моды на магнитную бурю с *K*p=5 SSC.

Анализ нескольких полей данных

До сих пор речь шла об анализе поля данных одной величины. Теперь предположим, что нас интересует пространственно-временная связь между двумя величинами. Число точек измерения и их местоположение могут быть абсолютно разными. То есть мы, например, можем анализировать поля TEC над Европой и над Северной Америкой или результаты глобальной модели на разных высотах. Единственное ограничение – измерения должны быть сделаны в одно и то же время. Пусть данные по величине 1 организованы в матрицу F размером $N \times Q$, а по величине 2 – в матрицу G размером $N \times R$. Формируем матрицу ковариаций $S = F^T G$ размера $Q \times R$ и вычисляем ее сингулярное разложение $S \rightarrow U \cdot \Sigma \cdot V^T$. Сингулярные вектора для F – это столбцы U, а сингулярные вектора для G – столбцы V^T. Временные главные компоненты рассчитываются как $A = FU$ и $B = = GV$, реконструкция исходных данных выполняется обычным способом: $F = AU^T$ и $G = BV^T$. Квадрат взаимной ковариации рассчитывается по формуле

$$SCF_i = \sigma_{ii}^2 \Big/ \sum_i \sigma_{ii}^2 .$$

Для представления и анализа результатов полезно построить карты гетерогенной корреляции. По определению – это вектор значений коэффициентов корреляции между k-й главной компонентой поля № 1 и сеточными значениями поля № 2. Построив карту квадратов корреляций, мы получим распределение локальной взаимной вариации, объясняемой k-й модой.

Чтобы правильно отобрать сингулярные вектора для взаимного анализа и построения карт гетерогенной корреляции, необходимо прибегнуть к методу Монте-Карло. Пусть мы имеем 40 лет месячных наблюдений за двумя величинами. Создаем искусственные наборы данных (surrogate test по западной номенклатуре), перемешивая данные в каждой точке наблюдений случайным образом, с тем чтобы нарушить хронологический по-

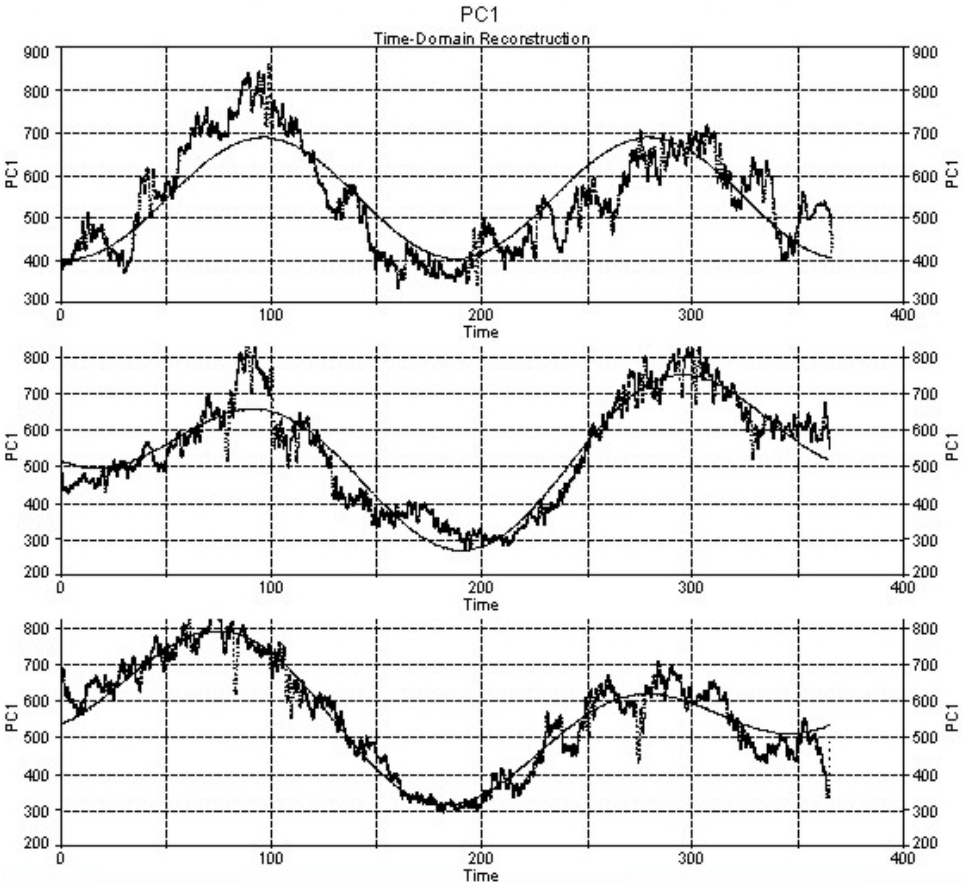

Рис. 4. Главные компоненты первой моды для 2000–2002 гг. (сверху вниз).

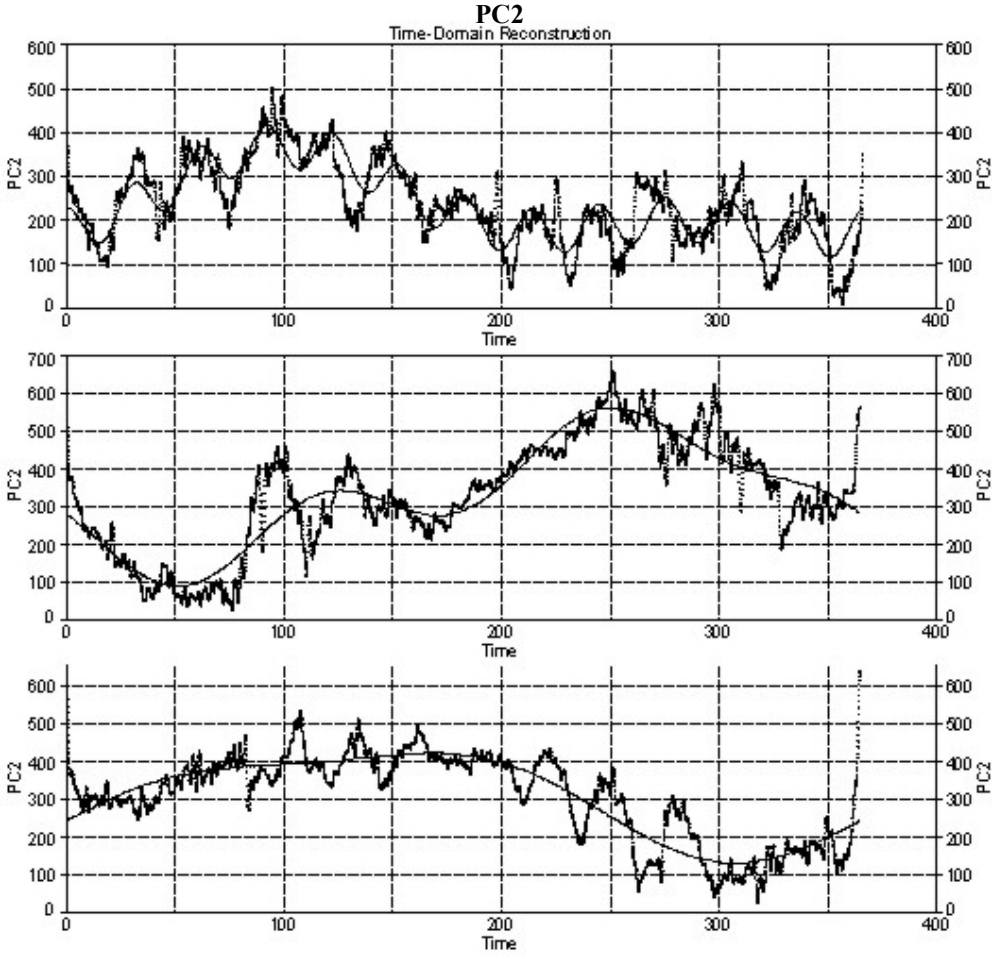

Рис. 5. Временные компоненты второй моды для 2000–2002 гг. (сверху вниз).

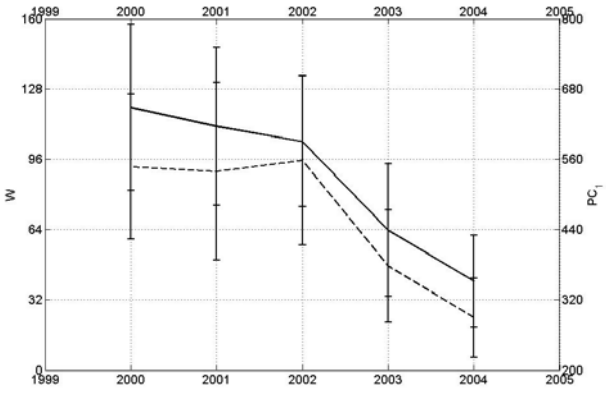

Рис. 6. Солнечная активность (левая ось Y) в числах Вольфа и амплитуда временной компоненты первой моды ТЕС за 2000–2004 гг.

рядок. Но при этом мы сохраняем распределение данных по сезонам. То есть перемешиваются года, но не месяцы. Январь продолжает соответствовать январю. Выполняется сингулярный анализ искусственных наборов данных (например, 100 реализаций), при этом мы запоминаем значения полного квадрата ковариации (сумма квадратов сингулярных значений) и квадрата ковариации *SCF* для каждой моды. Полученное на оригинальных данных значение *SCF* будет статистически значимым на 95 % уровне, если его превысило не более 5 % искусственных *SCF*.

Анализ данных полного электронного содержания в ионосфере

Для анализа были взяты данные полного электронного содержания за период 2000–2004 гг. Использовался аппарат комплексных ЕОФ. На рис. 1, 2 представлена пространственная структура первой и второй ЕОФ. Доля объясненной вариации для первой ЕОФ менялась от 62 % до 72 %, а второй ЕОФ – от 17 % до 22 %. Первая мода имеет выраженный сезонный и суточный ход (рис. 4), а ее амплитуда хорошо коррелирует с солнечной активностью (рис. 6). В пространственной картине второй ЕОФ выделяются приполярные зоны. При этом данная мода практически стационарна в магнитоспокойные периоды, резко меняясь во время магнитных бурь (рис. 3). Временная динамика моды меняется от года к году (рис. 5), коррелируя с уровнем магнитной активности (Ляхов А.Н., Хлыбов Е.С. // ДАН. 2006. Т. 409, N 6. С. 819–821.)

Для дальнейшего чтения…

В одной короткой лекции невозможно охватить все аспекты применения методов пространственно-временного анализа данных. К сожалению, отечественной современной литературы на эту тему практически не существует. Из зарубежной для дальнейшего ознакомления можно порекомендовать:

1. Analysis of Climate Variability. Applications of Statistical Techniques / Eds. H. von Storch and A. Navarra. Springer, 1995.

В книге подробно обсуждается метод ЕОФ, а также более развернутые методы – главных осциллирующих структур (Principal Oscillation Patterns), телесвязи (Teleconnection Patterns), Singular Spectrum Analysis, расчет карт пространственной корреляции.

2. Ghil M., Allen M.R., et al. Advanced Spectral Methods for climatic time series // Rev. Geophys. 2002. V. 40, N 1. P. 1-1–1-41. DOI 10.1029/2001RG000092.

Великолепный обзор по сингулярному спектральному анализу во всех вариантах и по методам ЕОФ. Содержит большое количество ссылок на оригинальные работы.

3. http://web.gfi.uib.no/~ngbnk/kurs/notes/course.html - Environmental Statistics for Climate Researches.

Интернет-ресурс по теме данной лекции и шире, содержит всю необходимую информацию, начиная с базовых понятий статистики.

Следующие три работы посвящены решению вопроса как быть, если данные известны с ошибкой и сама ошибка может быть оценена.

4. Denoeux T., Masson M.H. Principal component analysis of fuzzy data using autoassociative neural networks // IEEE Transactions on Fuzzy Systems. 2004. V. 12. P. 336–349.

5. D'Urso P., Giordani P. A least squares approach to principal component analysis for interval valued data // Chemometrics and Intelligent Laboratory Systems. 2004. V. 70. P. 179–192.

6. Giordani P., Kiers H.A.L. Principal Component Analysis of symmetric fuzzy data // Computational Statistics and Data Analysis. 2004. V. 45. P. 519–548.