УДК 620.678:539.371:518.12
А.М. Сулейманов

АНАЛИЗ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ МЕТОДОМ ГЛАВНЫХ КОМПОНЕНТ

При изучении сложных физических и химических систем можно выделить два принципиально разных подхода:

- детальное изучение процессов, происходящих в системе, и построение содержательных моделей (белые модели), обычно в виде систем дифференциальных или интегро-дифференциальных уравнений, далее - применение специальных пакетов или методов для их решения;

- гибкий и достаточно надежный способ анализа данных, основанный на многофакторном формальном моделировании (черные модели).

Каждый из этих подходов имеет свои плюсы и минусы, но в любом случае построенная модель, сколь сложна она ни будет, - всегда некоторое приближение к реальности. Однако хорошая модель является эффективным инструментом для анализа структуры данных. Формально-математический подход особенно эффективен в случаях, когда непонятно, как строить содержательную модель, либо на ее построение и дальнейшие вычисления требуются чрезмерные усилия.

В основе многофакторного анализа данных лежат проекционные математические методы. Эти методы позволяют выделить в больших массивах данных скрытые (латентные) переменные и анализировать связи, существующие в изучаемой системе. На западе такой подход, получивший название Chemometrics (хемометрика), бурно развивается: регулярно проводятся конференции, издаются журналы, такие как *Journal of Chemometrics* и *Chemometrics and Intelligent Laboratory Systems*. Важное место в методах анализа многомерных данных занимает метод главных компонент (МГК) - Principal Component Analysis (PCA). Идея метода главных компонент была сформулирована английским математиком Карлом Пирсоном [1] в 1901 году, и с тех пор плодотворно используется для анализа внутренних закономерностей в больших массивах. Центральная концепция такого подхода это понятие главного компонента. Так называют специальный тип переменной - латентную переменную, которая не может быть явно объявлена и непосредственно измерена. В математическом смысле латентная переменная является линейной комбинацией исходных переменных [2], которая может быть формально определена как собственный вектор ковариационной матрицы данных. Главные компоненты [рис. 1] показывают скрытые систематические связи, присущие исходному набору данных. При этом новая модель имеет, как правило, существенно меньшее количество переменных, в силу чего такой подход и может интерпретироваться как проекционный, когда исходные данные проецируются на гиперплоскость [рис. 2] размерности меньшей, нежели исходное пространство.

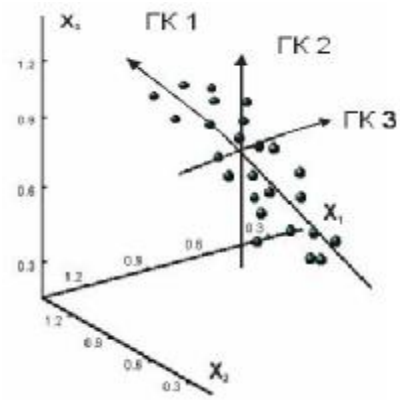

Рис.1. Поиск главных компонент модели

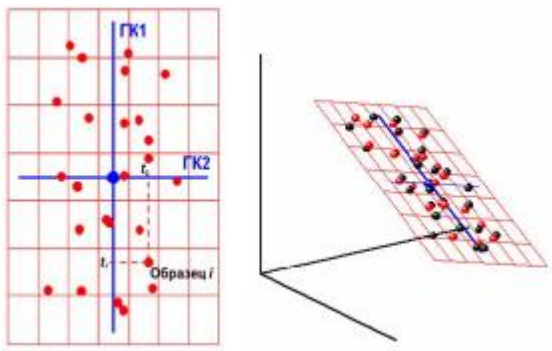

Рис.2. Проекция данных на плоскость главных компонент

При построении моделей с помощью проекционных методов основными являются два момента. Первый - это калибровка [2], т.е. создание модели [рис.3] исходных данных. Центральный момент здесь - это определение скрытой

размерности системы, т.е. количество латентных переменных. При таком выборе необходимо учитывать, что слишком малое количество переменных плохо опишет данные и оставит существенную часть информации вне модели. С другой стороны, слишком подробное описание модели, с помощью большого количества главных компонент, будет моделировать не только систематические связи в системе, но также и различные ошибки, неизбежно присутствующие в любых данных, что приведет затем к ошибкам при прогнозировании. Поэтому вторым важным моментом при построении модели является ее проверка, позволяющая исследовать качество предлагаемой модели. Отдельно необходимо изучать вопрос пригодности модели для последующего прогнозирования [рис.4]. При этом необходимо отметить, что матрицы входных (X) и выходных (Y) параметров могут состоять практически из неограниченного количества переменных. Также необходимо иметь в виду и другие стадии моделирования, такие как предварительная обработка данных и выявление выбросов. Хотя данный подход и называется формально-математическим, для эффективного анализа данных необходимо не только знание основ билинейного моделирования, но также очень важно и понимание сущности изучаемой системы.

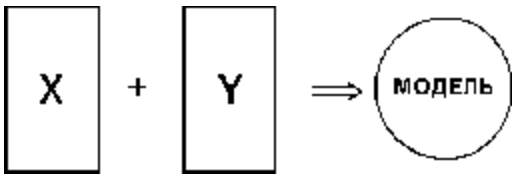

Рис.3. Калибровка. Построение регрессионной модели по известным данным X и Y

Рис.4. Использование многомерной регрессионной модели для предсказания новых значений Y

Основываясь на вышеописанном подходе, по полученным экспериментальным данным [3], используя пакет программ Unscrambler ® фирмы CAMO, мы [4] построили модель для анализа зависимости эксплуатационных показателей (ЭП) ПВХ-профилей от состава смеси. Исследовалось влияние 2-х основных типов добавок, модификатора ударной прочности и стабилизатора. Исходный массив данных состоял из 15 образцов [табл], из них 11 образцов использовались для построения модели (калибровочные образцы) и 4 образца - в качестве тестовых . В качестве предикторов были выбраны массовые части добавок: два типа модификатора ударной прочности МУП (FM 22 и дакрилан); два типа стабилизатора (интерстаб 3009, нафтомикс GMX). Т.е. 4 переменные в матрице X. В качестве откликов использовались следующие ЭП: белизна (%); предел прочности при растяжении (МПа); ударная прочность (кДж/м^2); показатель текучести (г/10мин); т.е. 4 переменные в матрице Y.

Естественно, строя МГК-модель для анализа данных такой относительно простой системы, мы не ставили задачи открыть что-то новое для материаловедения, а хотели просто показать, что данный подход дает всеобъемлющее представление о структуре данных, которое можно охватить одним взглядом. Например, модель (график нагрузок [рис 5]) показывает, что:

- МУП (FM 22 и дакрилан) располагаются рядом, т.е. они действуют на отклики одинаково;
- то же самое можно сказать и о стабилизаторах (интерснаб и нафтомикс);
- между МУП и стабилизаторами - отрицательная корреляция;
- между прочностью на растяжение и текучестью - также отрицательная корреляция;
- расположение белизны на нуле говорит о том, что на этот ЭП эти добавки не оказывают никакого действия.

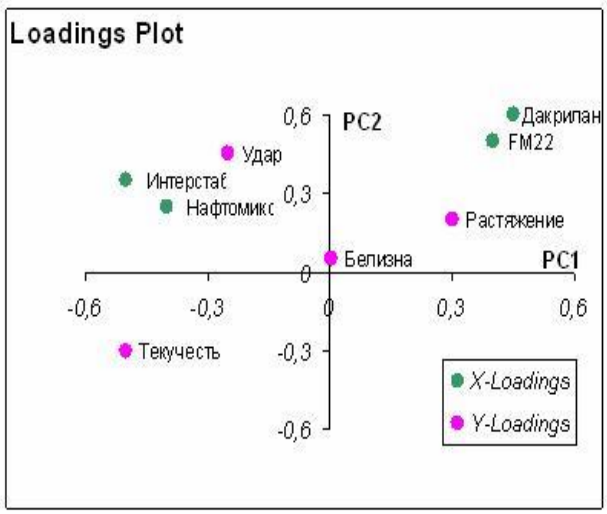

Рис.5. Графики нагрузок для ГК1 и ГК2

Таблица

№		X							Y			
		МУП		Стабилизатор								
		FM22	Дакрилан	Интерстаб 3009	Нафтомикс GMX	Наполнитель	Пластификатор	Пигмент	Белизна	Растяжение	Удар	Текучесть
1	калибровочные	7	0	5,5	0	6	0,1	5	82,5	43,5	30	0,2
2		6	0	5,5	0	6	0,1	5	85,5	45,5	26	0,15
3		5	0	4,2	0	6	0,1	5	85	44	28	0,17
4		0	6	5	0	6	0,1	5	84	55	28	0,008
5		0	7	5,5	0	6	0,1	5	83,5	53	26	0,3
6		0	8	6	0	6	0,1	5	86	43	26	0,2
7		6	0	0	5,5	6	0,1	5	85,5	45	32	0,15
8		6	0	0	6	6	0,1	5	84,5	46	30	0,12
9		7	0	0	5,3	6	0,1	5	86,5	48	24	0,13
10		7	0	0	3	6	0,1	5	84	44,5	31	0,17
11		5	0	0	6,5	6	0,1	5	83,5	42	28	0,2
12	тестовые	6	0	0	4	6	0,1	5	83,5	55	26	0,2
13		0	6	4	0	6	0,1	5	84,5	48	30	0,15
14		5	0	0	4,5	6	0,1	5	85	42,5	28	0,25
15		0	6,3	5,5	0	6	0,1	5	86,5	43,5	31	0,1

ЛИТЕРАТУРА

1. Pirson K. On lines and planes of closest fit to systems of points in space. Phil.Mag. (6), 2, 559-572, 1901.
2. Esbensen K.H. Multivariate Data Analysis - In Practice 4-th Ed., CAMO, (2000), 598p (ISBN 82-993330-2-4).
3. Абдрахманова Л.А., Шакуров Ф.Г., Сулейманов А.М.., Хозин В.Г. Поведение жестких поливинилхлоридных профилей при естественном и ускоренном старении //Деструкция и стабилизация полимеров. Тезисы докладов 9-ой конференции. РАН М., 2001. - С. 3-4.
4. Сулейманов А.М., Абдрахманова Л.А., РодионоваО.Е., Померанцев А.Л. Формально-математический подход к анализу многомерных массивов данных при исследовании свойств полимеров// Материалы научных трудов Вторых Воскресенских чтений “Полимеры в строительстве”. Казань, 2004. - С.76.