Федеральная служба по гидрометеорологии
и мониторингу окружающей среды

**Всероссийский НИИ
сельскохозяйственной метеорологии**

Оценка средней районной урожайности зерновых культур на основе наземной и спутниковой информации с использованием метода главных компонент.

А.Д. Клещенко, О.В. Савицкая, С.А. Косякин.

- Существующий метод расчета средней районной урожайности зерновых культур на основе спутниковой и наземной информации.

- Применение метода главных компонент для оценки средней районной урожайности зерновых культур.

Входные данные:

- Статистическая информация: **средне-районная** урожайность (Федеральная служба государственной статистики, база данных показателей муниципальных образований);
- Спутниковая информация: новый информационный продукт **IKI MODIS LAI**, **NDVI**, **VCI** (ИКИ, сервис ВЕГА-PRO).

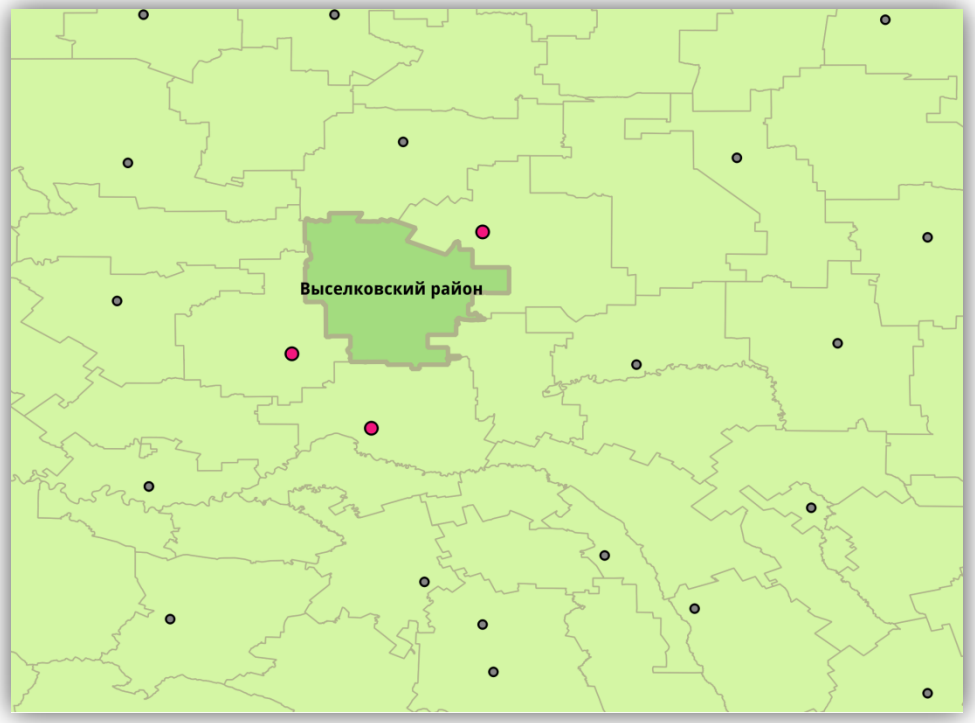
$$VCI_i = \frac{100 * (NDVI_i - NDVI_{min})}{NDVI_{max} - NDVI_{min}}$$, где $NDVI_i$ - значение NDVI для даты j;

$NDVI_{max}$ - максимальное значение NDVI внутри всего набора данных;

$NDVI_{min}$ - минимальное значение NDVI внутри всего набора данных.

- Наземная информация: декадные и срочные агрометеорологические данные по станциям.

– Станция расположена внутри района;
– Станция внутри района отсутствует, расчет осуществляется по данным трех ближайших станций.



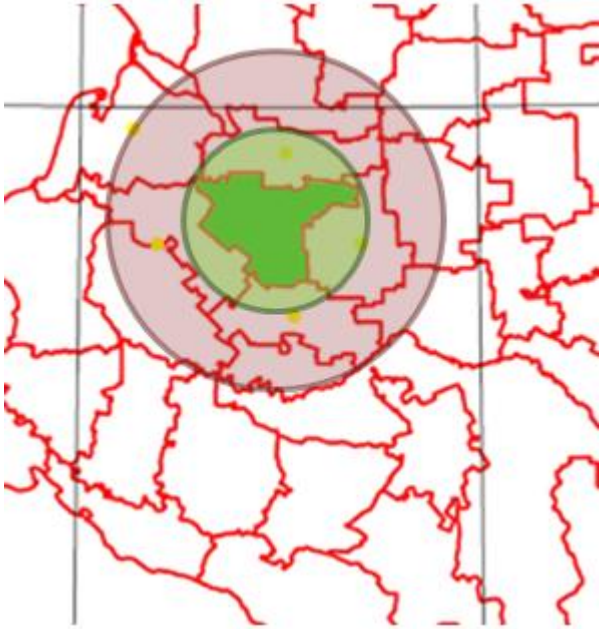
Ближайшая точка вносит больший вклад в интерполируемое значение, чем более удаленная.

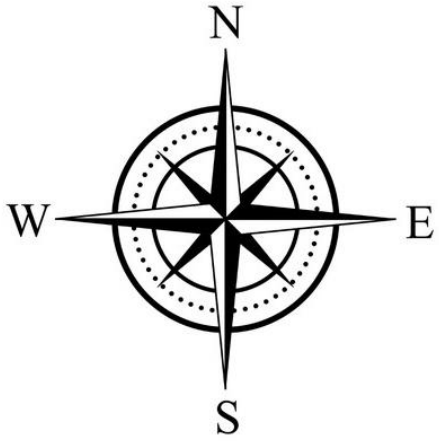
$$E = \frac{\sum_{i=1}^{n} w_i E_i}{\sum_{i=1}^{n} w_i}$$

$$w_i = \frac{1}{r_i^2}$$

где E – рассчитываемое средневзвешенное значение метеорологического параметра;
E_i - значения метеорологического параметра в ближайших точках, попавших в заданную окрестность;
w_i - рассчитываемый вес i-ой точки – обратная функция расстояния;
r_i - расстояние от точки интерполяции до i-ой точки.

Критерий выбора ближайших агрометеорологичеких станций



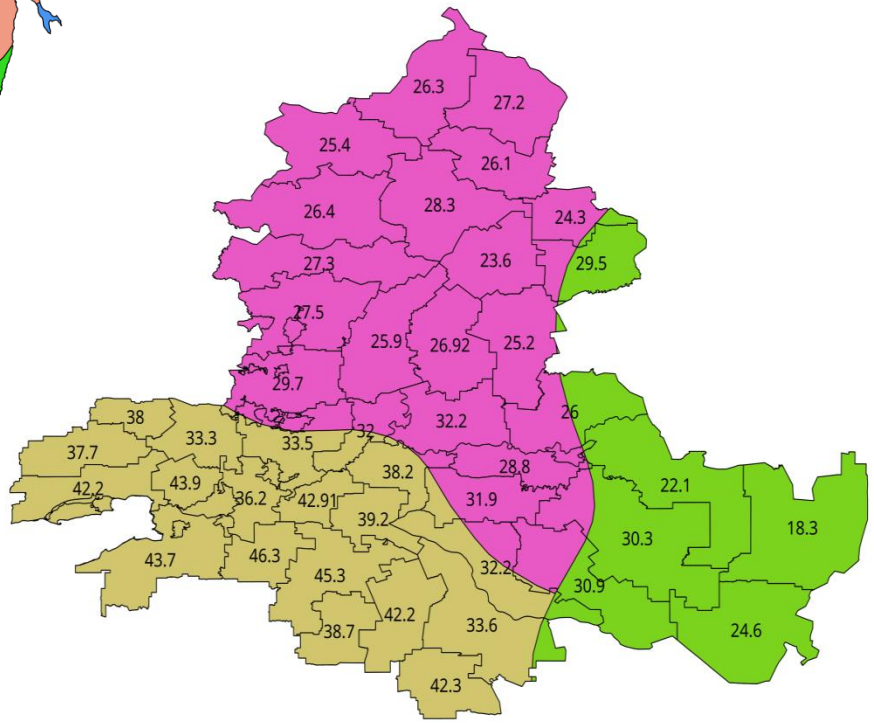


Первый этап– выбор оптимальной удаленности станций.
Вычисляются центы районов, граничащих с анализируемым. Максимальное расстояние от самого удаленного центра соседнего района уменьшается вдвое, если же нужное количество станций не найдено в получившемся радиусе(на рис. отмечен зеленым кругом), то берутся станции из полученного изначально радиуса (на рис. бордовый круг).

Второй этап – расположение станции по сторонам света.

Максимальный радиус не превышает 80 км.

Дифференциация территории на зоны на основе карты агроклиматического районирования территории, разработанной Д.И. Шашко



– Временной диапазон 6 лет: с 2012 по 2017 гг.

– Данные по районам объединялись в группы для увеличения объема выборки.

– Центрирование и нормирование данных:

нахождение среднего

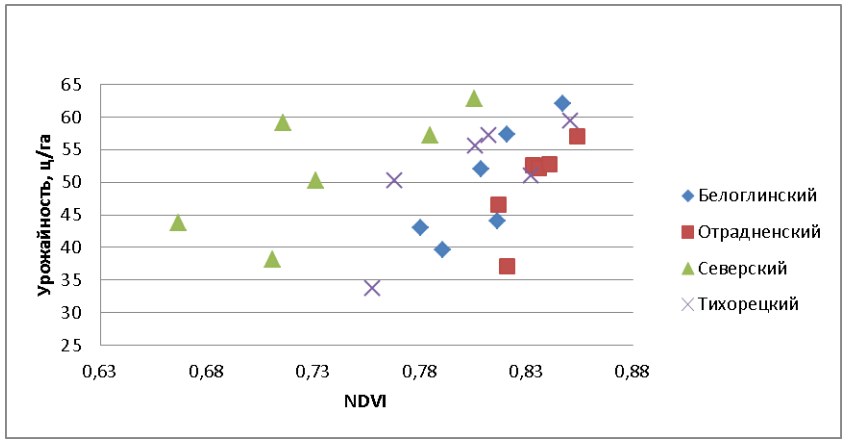
$$\bar{V} = \frac{\sum_{i=1}^{n} V_i}{n}$$

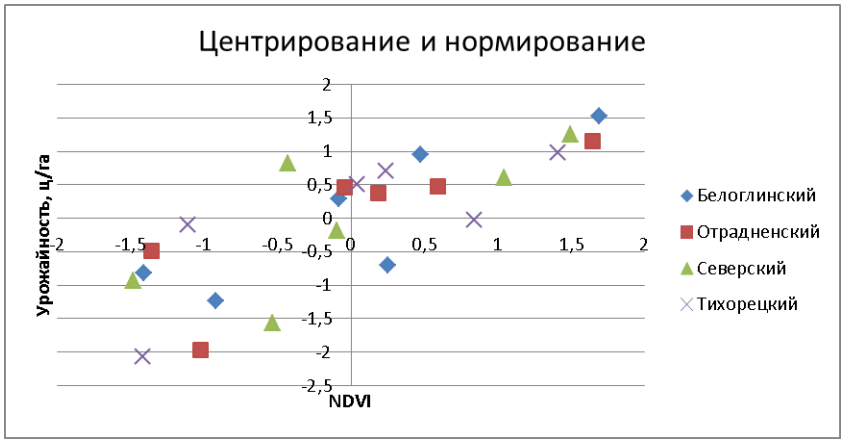
разность между исходными числами и их средним

$$X_i = V_i - \bar{V}$$

нормирование, путем деления на среднеквадратическое отклонение (сигму, σ)

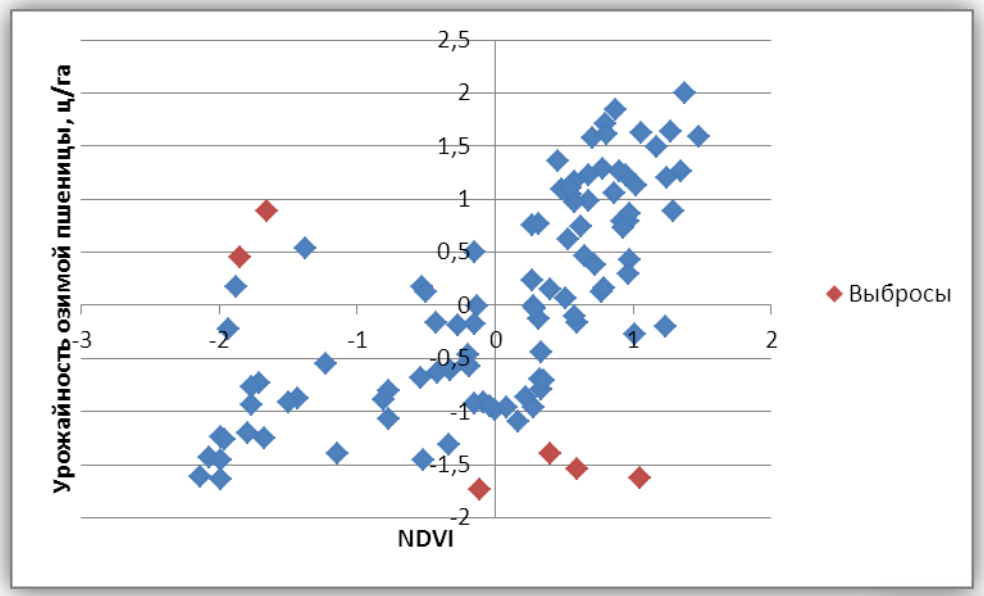
$$x_i = X_i / \sigma \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n} (V_i - \bar{V})^2}{n}}$$





Выбросы, Оренбургская область, 3 декада мая

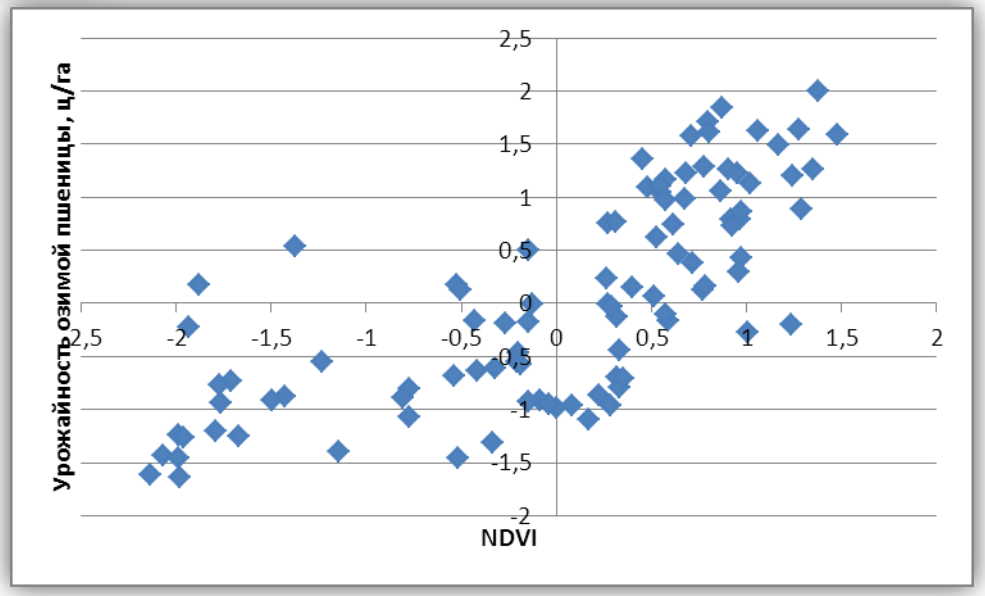
R=0,62



Стандартизованные остатки выходят за пределы диапазона от -2 до 2

Стандартизированные остатки – это остатки, деленные на собственное среднеквадратическое отклонение

R=0,74



Коэффициенты корреляции между спутниковыми индексами и средней районной урожайностью озимой пшеницы

Ростовская область

Месяц	Декада	Группа 1		Группа 2		Группа 3		Группа 4	
		NDVI	LAI	NDVI	LAI	NDVI	LAI	NDVI	LAI
Май	1	0,78	0,87	0,80	0,88	0,76	0,79	0,82	0,88
Май	2	0,90	0,93	0,77	0,84	0,85	0,90	0,86	0,92
Май	3	0,92	0,96	0,91	0,90	0,89	0,96	0,85	0,94
Июнь	1	0,81	0,93	0,84	0,89	0,82	0,89	0,76	0,86

Волгоградская область

Месяц	Декада	Группа 1		Группа 2	
		NDVI	LAI	NDVI	LAI
Май	1	0,71	0,77	0,78	0,83
Май	2	0,89	0,91	0,86	0,86
Май	3	0,89	0,92	0,86	0,86
Июнь	1	0,87	0,89	0,81	0,84

Уточненные регрессионные модели

для районов Ростовской области

Месяц	Декада	Коэффициенты уравнений за период 2012-2016 гг.			Коэффициенты уравнений за период 2012-2017 гг.		
		a	b	c	a	b	c
Май	2	0	-0,44	0,51	0	-0,37	0,59
Май	3	0	-0,36	0,79	0	-0,31	0,80
Июнь	1	0	-	0,71	0	-	0,81

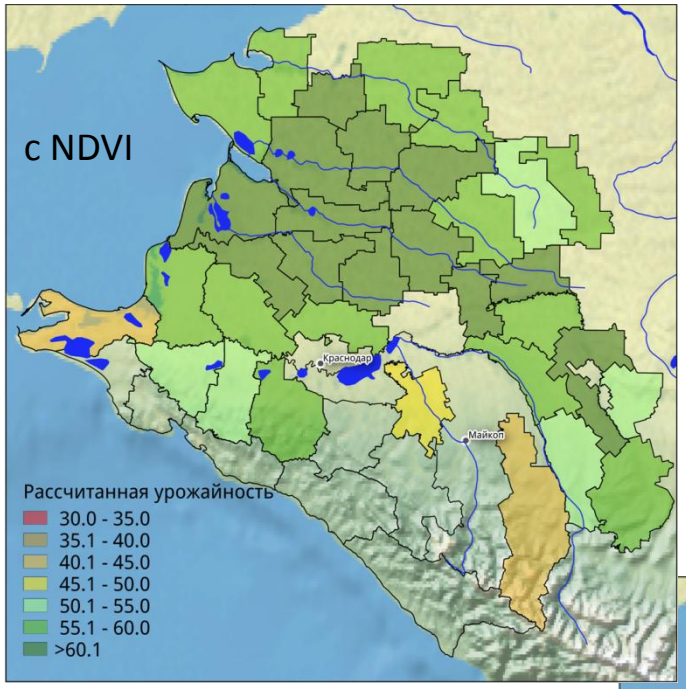
a – свободный член
b – коэффициент при дефиците влажности воздуха
c – коэффициент при NDVI

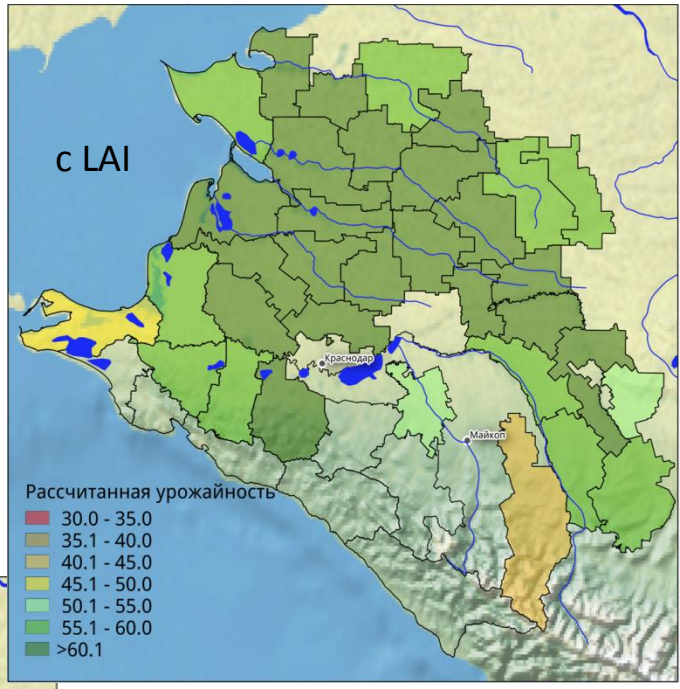
Месяц	Декада	Коэффициенты уравнений за период 2012-2016 гг.			Коэффициенты уравнений за период 2012-2017 гг.		
		a	b	c	a	b	c
Май	2	0	-0,43	0,56	0	-0,33	0,66
Май	3	0	-0,30	0,86	0	-0,22	0,87
Июнь	1	0	-	0,90	0	-	0,93

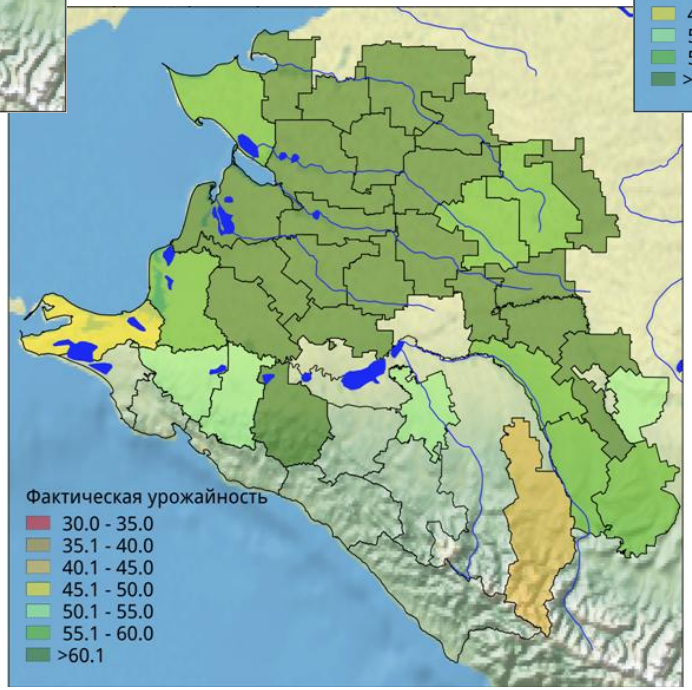
a – свободный член
b – коэффициент при дефиците влажности воздуха
c – коэффициент при LAI

Сравнение рассчитанных и фактических районных урожайностей озимой пшеницы, Краснодарский край

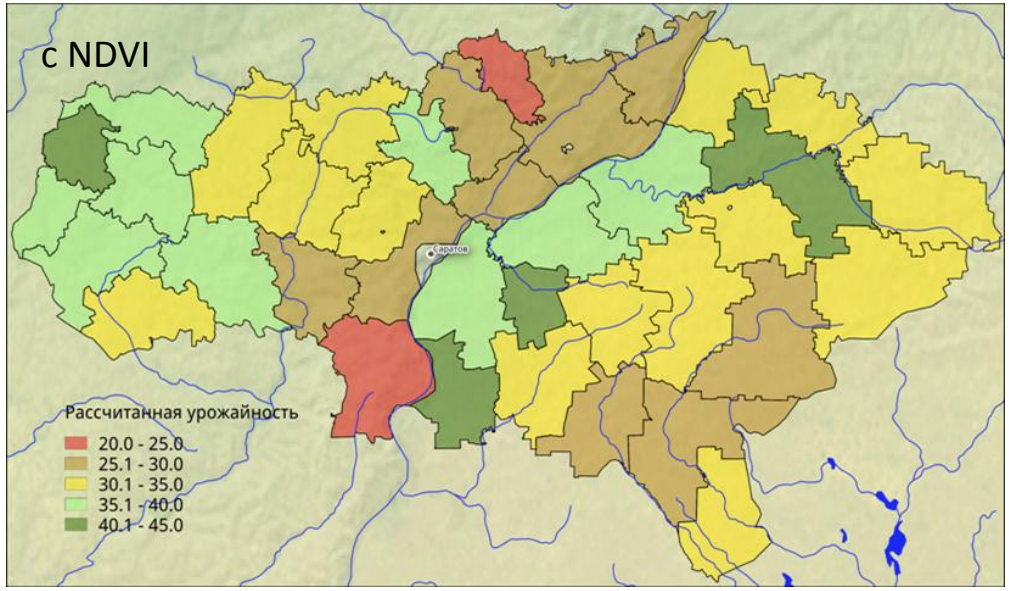
3 декада мая 2017 г.





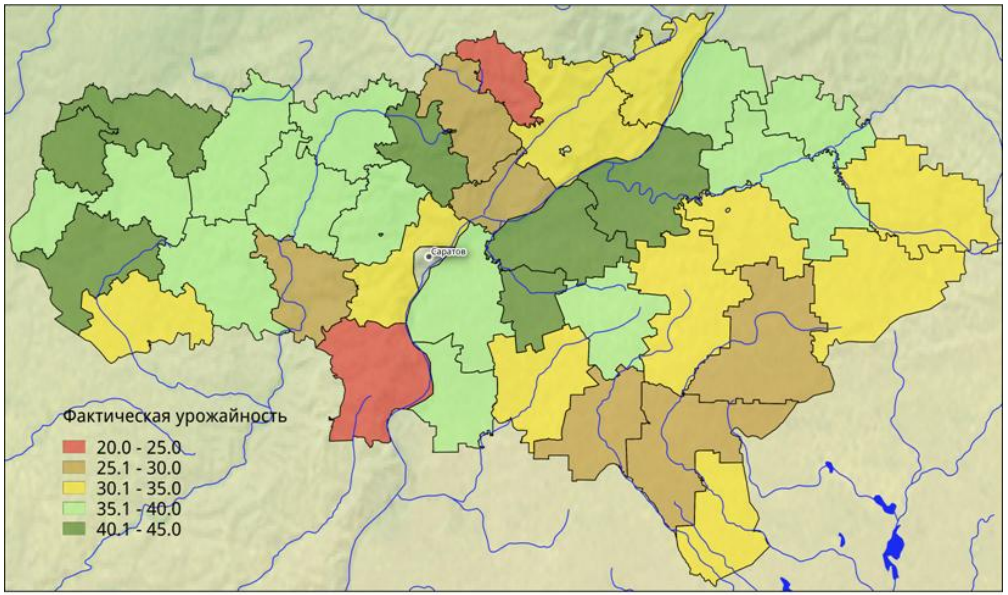


Сравнение рассчитанных и фактических районных урожайностей озимой пшеницы, Саратовская область



2017 июнь 1 декада

Относительная ошибка, %	Количество районов	Процент районов, %
менее 5	22 из 38	58
от 5 до 10	11 из 38	29
более 10	5 из 38	13



Корреляционная матрица, Ростовская область, 2 декада мая

	ndvi	lai	T	осадки	дефицит	урожайность
ndvi	1					
lai	0,94	1				
T	-0,64	-0,65	1			
осадки	0,38	0,40	-0,55	1		
дефицит	-0,69	-0,66	0,95	-0,72	1	
урожайность	0,68	0,79	-0,65	0,76	-0,73	1

- Включение в регрессионную модель мультиколлинеарных факторов не совсем корректно.
- В этом случае оценки параметров регрессии не надежны, отсюда следует, что модель не пригодна для анализа и прогнозирования.

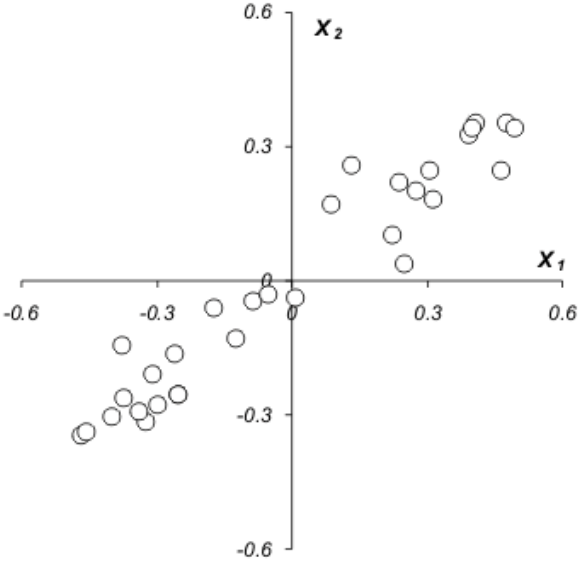
Метод главных компонент ориентирован на выделение в многомерном пространстве группы тесно коррелирующих между собой переменных и замене их без потери информативности главными компонентами, между которыми корреляция отсутствует

Преимущество метода главных компонент:

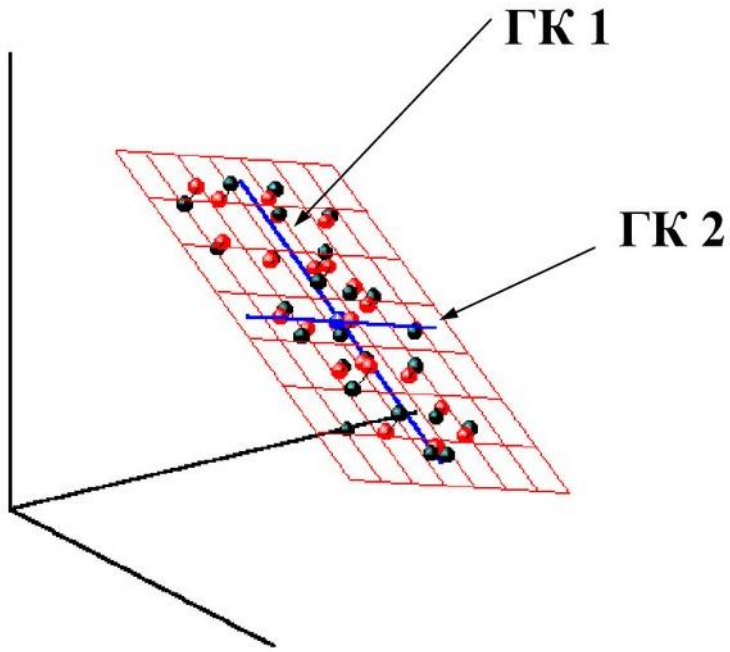
- Избаление от мультиколлинеарности
- Некоррелируемость главных компонент между собой
- Эффективный способ снижения размерности данных, позволяет сохранить максимум информации в минимальном количестве переменных

Графическое представление метода главных компонент

	x_1	x_2
1	0.407	0.353
2	0.475	0.355
3	-0.088	-0.045
4	0.394	0.325
5	0.274	0.202
6	0.131	0.258
7	-0.053	-0.031
8	-0.124	-0.128
9	-0.469	-0.344
10	0.088	0.171
11	-0.261	-0.162
12	0.401	0.341
13	-0.376	-0.143
14	-0.251	-0.255
15	-0.325	-0.316
16	0.464	0.248
17	-0.310	-0.207
18	0.307	0.247
19	-0.399	-0.303
20	-0.253	-0.253
21	-0.341	-0.291



- Выбирается направление, которому соответствует максимальная дисперсия, т.е. наибольшая дифференциация, разброс объектов. Это первая главная компонента (ГК1);
- Затем выбирается еще одно направление (ГК2), ортогональное к первому, так чтобы описать оставшееся изменение в данных и т.д.
- Для каждой следующей компоненты дисперсия убывает, а последняя компонента будет иметь наименьшую дисперсию



- переменные трансформируются в новые не коррелирующие друг с другом главные компоненты;

- главные компоненты являются линейными комбинациями исходных переменных;

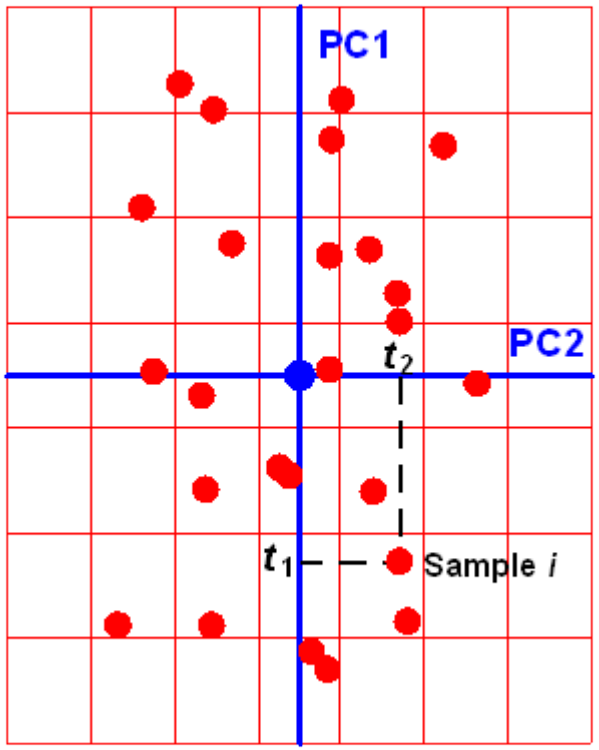
$$T_{iA} = C_1 y_{i1} + C_2 y_{i2} + \cdots + C_j y_{ij} + \cdots + C_p y_{ip}$$

$$T_{(i+1)A} = C_1 y_{(i+11)} + C_2 y_{(i+1)2} + \cdots + C_j y_{(i+1)j} + \cdots + C_p y_{(i+1)p}$$

$$\cdots\cdots$$
$$\cdots\cdots$$
$$\cdots\cdots$$

$$T_{IA} = C_1 y_{i1} + C_2 y_{i2} + \cdots + C_j y_{ij} + \cdots + C_p y_{ip}$$

где p – количество переменных;
A – количество компонент, изменяется от 1 до p;
i – изменяется от 1 до I.
I – количество наблюдений;



$$\mathbf{T}= \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1a} & \dots & t_{1A} \\ t_{21} & t_{22} & \dots & t_{2a} & \dots & t_{2A} \\ & & & & & \\ t_{i1} & t_{i2} & \dots & t_{ia} & \dots & t_{iA} \\ & & & & & \\ t_{I1} & t_{I2} & \dots & t_{Ia} & \dots & t_{IA} \end{bmatrix}$$

Матрица **T** дает проекции исходных переменных на подпространство главных компонент.
Строки матрицы **T** соответствуют количеству наблюдений.
Столбцы матрицы **T** – ортогональны и представляют проекции всех переменных на одну новую координатную ось.

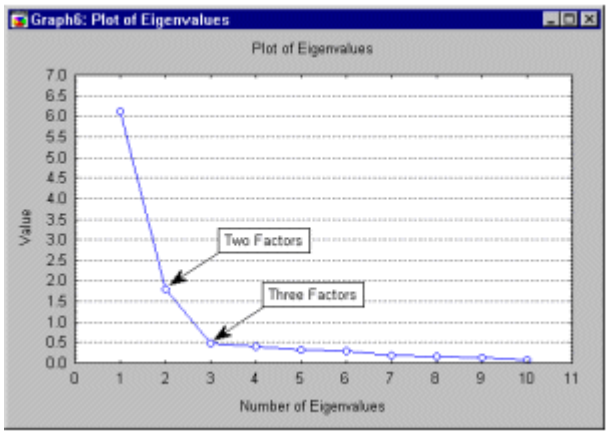
- Если число главных компонент слишком мало, то описание данных будет не полным.

- Избыточное число главных компонент приводит к переоценке, т.е. к ситуации, когда моделируется шум, а не содержательная информация.

Критерии выбора:

- Критерий Кайзера (Kaiser, 1960), отбираются только факторы, с собственными значениями, большими *1*.

- Критерий каменистой осыпи (Cattell, 1966), графический метод. Находится место на графике, где убывание собственных значений слева направо максимально замедляется.



Вращение факторов

Компоненты **легко интерпретировать** если: каждая исходная переменная коррелирует только с одной компонентой; нагрузки (loadings) близки либо 1/-1, либо 0

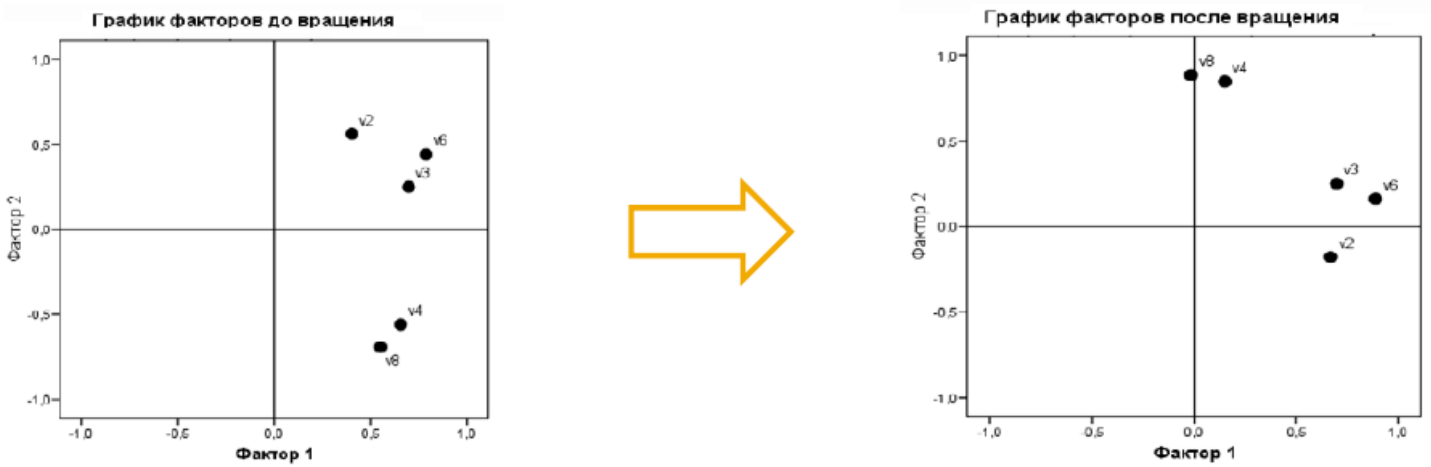
Сложно интерпретировать если: среди нагрузок (loadings) много невысоких значений; некоторые переменные почти одинаково коррелируют с несколькими компонентами

Целью вращения является, получение простой структуры, при которой каждая переменная коррелирует не более чем с одной компонентой.

Варимакс (*Varimax*) - наиболее распространенный метод вращения, при котором при сохранении ортогональности факторов минимизируется число переменных с высокой факторной нагрузкой.

Метод варимакс максимизирует дисперсию квадратов нагрузок для каждого фактора, что приводит к увеличению больших и уменьшению малых значений факторных нагрузок.

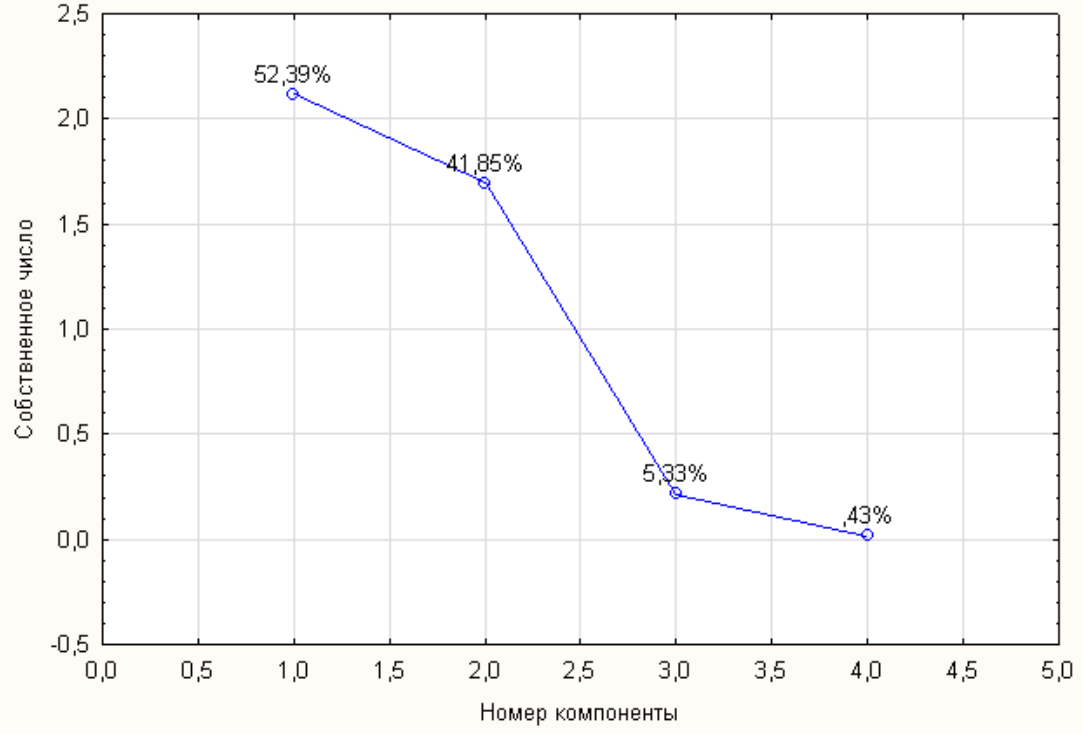
Факторные нагрузки



После поворота осей, переменные оказываются вблизи осей, что соответствует максимальной нагрузке каждой переменной только по одному фактору

Компо-нента	Собствен-ные числа	% общей дисперсии	Кумулят. соб. числа.	Кумулят. % общ. дисп.
1	2,12	52,39	2,12	52,39
2	1,69	41,85	3,81	94,25
3	0,22	5,33	4,03	99,57
4	0,02	0,43	4,04	100,00

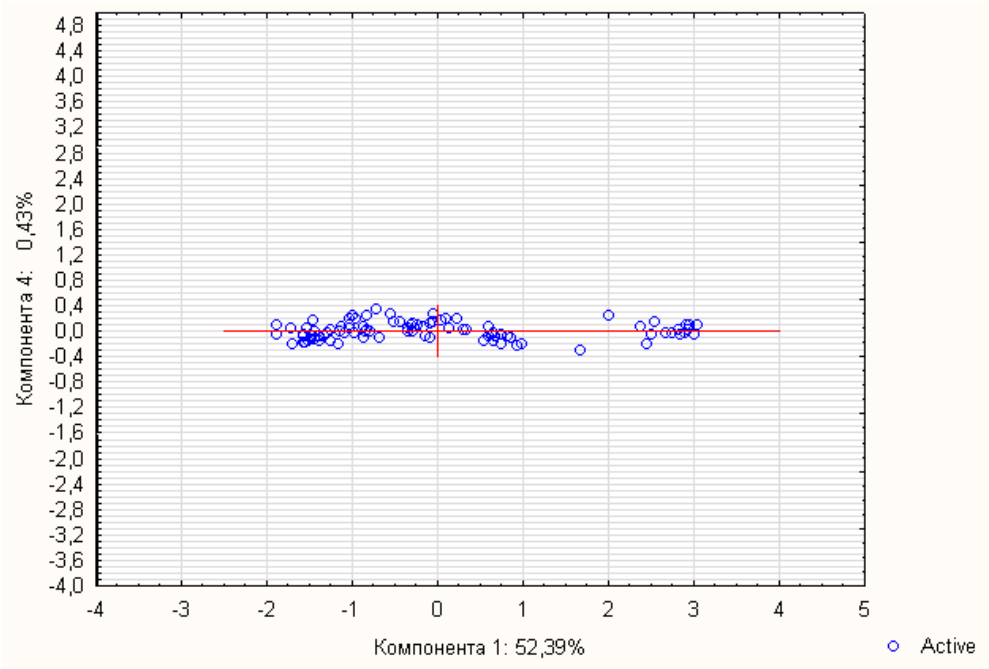
NDVI, LAI, дефицит, осадки



Коэффициенты корреляции переменных с главными компонентами, Ростовская область, 1 декада мая

Переменная	Компонента 1	Компонента 2	Компонента 3	Компонента 4
ndvi	0,94	-0,32	0,01	0,09
lai	0,89	-0,44	0,10	-0,09
дефицит	0,28	0,91	0,32	0,01
осадки	-0,58	-0,75	0,32	0,02

Проекции точек на факторную плоскость



Месяц	Декада	Коэффициенты уравнений регрессии за период 2012-2017 гг.			R
		св. член	1 компонента	2 компонента	
Май	1	0	0,33	-0,56	0,87
Май	2	0	0,52	-	0,90
Май	3	0	0,66	-0,06	0,95
Июнь	1	0	0,67	-	0,91

Месяц	Декада	1 компонента	2 компонента
Май	1	NDVI, LAI	осадки дефицит
Май	2	NDVI, LAI дефицит	-
Май	3	NDVI, LAI	дефицит
Июнь	1	NDVI, LAI	-

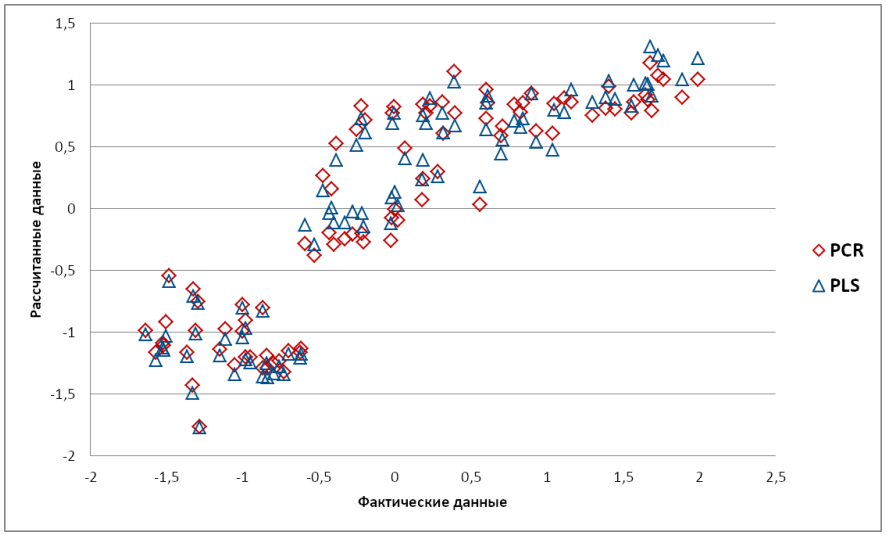
Ростовская область, Рассчитанная урожайность с 2012 по 2017 гг.

Группа	Метод	Относительная ошибка, %			
		1 декада мая	2 декада мая	3 декада мая	1 декада июня
1	Регрессия	14,8	12,1	9,9	14,8
1	МГК	12,2	11,5	8,0	11,0
2	Регрессия	23,2	23,2	11,0	15,3
2	МГК	19,2	18,6	11,5	13,0
3	Регрессия	10,5	9,9	8,7	11,3
3	МГК	10,5	9,0	6,5	8,5
4	Регрессия	14,1	12,8	12,4	14,6
4	МГК	12,9	10,4	10,4	12,0

Будущие исследования

РЛС – регрессия на латентные структуры.
При построении проекционной модели учитывается связь между x и y. Критерием является моделирование той структуры (информации) в X, которая имеет корреляцию с Y.

R^2 (МГК) – 0,75
R^2 (РЛС) – 0,79



MATLAB Neural Network Toolbox
Входные данные: главные компоненты

Метод	Относительная ошибка, %			
	1 декада мая	2 декада мая	3 декада мая	1 декада июня
Регрессия	14,8	12,1	9,9	14,8
МГК	12,2	11,5	8,0	11,0
Нейронная сеть	8,4	5,9	6,0	8,3

СПАСИБО ЗА ВНИМАНИЕ