Всероссийский НИИ
сельскохозяйственной метеорологии

Построение проекционных моделей для оценки ожидаемой урожайности озимой пшеницы на основе спутниковой и метеорологической информации.

А.Д. Клещенко, О.В. Савицкая

Основные разделы доклада

– Применение метода главных компонент для оценки средней районной урожайности озимой пшеницы на основе спутниковой и наземной информации для территории Северо-Кавказского УГМС.

– Применение метода проекции на латентные структуры для оценки средней районной урожайности озимой пшеницы на основе спутниковой и наземной информации для территории Центрально-Черноземного УГМС.

Входная информация

– Спутниковые индексы: **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;

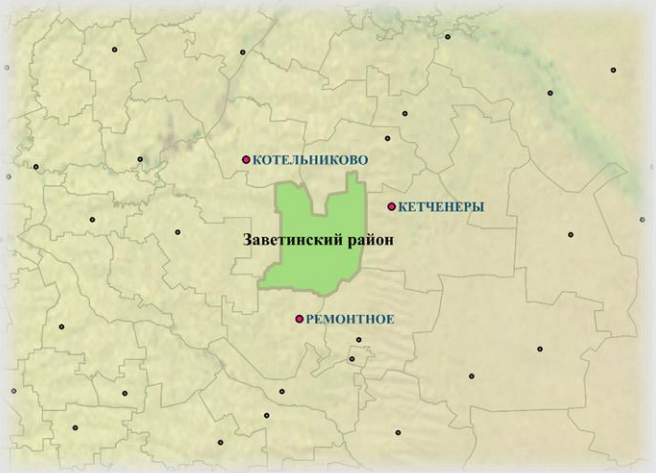
$NDVI_{min}$ - минимальное значение NDVI внутри всего набора данных.

– Наземная метеорологическая информация получена по данным наблюдений на гидрометеорологических станциях Росгидромета. Параметры: средняя температура воздуха за декаду и за 3 декады; сумма осадков за декаду и за 3 декады; средний дефицит влажности воздуха за декаду и за 3 декады; ГТК за месяц.

– Статистическая информация: **средняя районная** урожайность (Федеральная служба государственной статистики, база данных показателей муниципальных образований);

Метод обратных взвешенных квадратов расстояний (Ю.В. Ткачева, 2018 г.)

- Получение метеорологической информации для районов, в которых станции отсутствовали.
- Идея метода: ближайшая точка вносит больший вклад в интерполируемое значение, чем более удаленная.



$$E = \frac{\sum_{i=1}^{n} w_i E_i}{\sum_{i=1}^{n} w_i} \qquad w_i = \frac{1}{r_i^2}$$

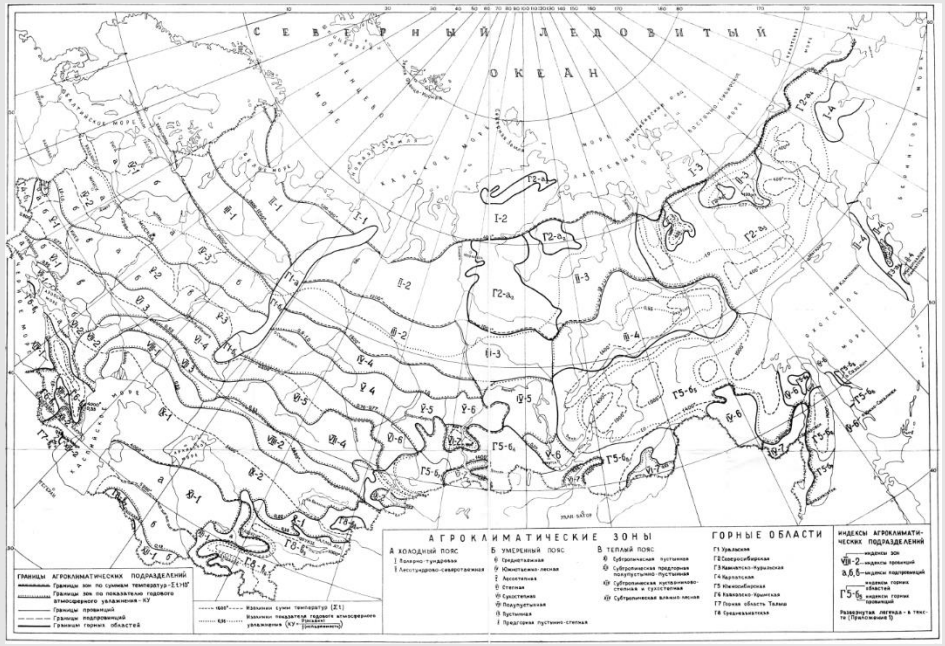
где E – рассчитываемое средневзвешенное значение метеорологического параметра;

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.

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

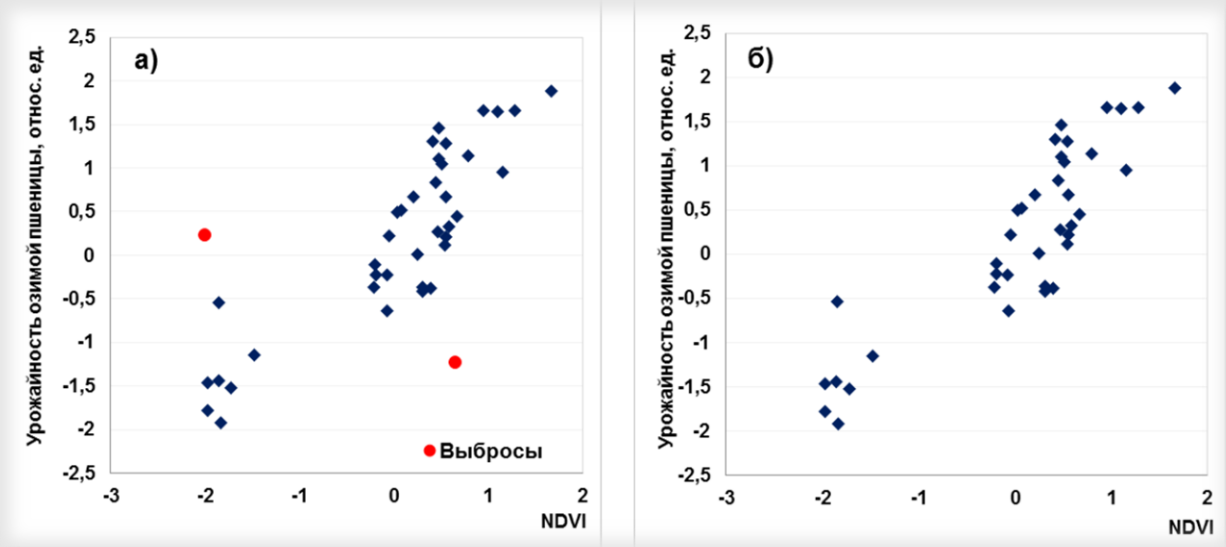


Предварительная обработка данных

- Период исследования: с 2012 по 2021 гг.
- Данные центрируются (вычитается среднее) и нормируются (деление на среднеквадратическое отклонение):
- Обнаружение статистических выбросов

$$x_i = X_i / \sigma$$

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



Стандартизованные остатки выходят за пределы диапазона от -2 до 2
Стандартизированные остатки – это остатки, деленные на собственное среднеквадратическое отклонение

Коэффициенты корреляции между исходными параметрами и средней районной урожайностью озимой пшеницы

Ставропольский край

Группа	Длина ряда	Месяц	Декада	Метеорологическая информация							Спутниковая информация	
				T	P	D	T3	P3	D3	GTK	NDVI	VCI
1	96	май	1	-0,82	0,69	-0,86	-0,78	0,53	-0,90	0,60	0,64	0,65
		май	2	-0,67	0,54	-0,79	-0,88	0,65	-0,95	0,69	0,81	0,81
		май	3	-0,43	0,17	-0,20	-0,80	0,63	-0,87	0,66	0,80	0,83
		июнь	1	-0,41	-0,11	-0,11	-0,64	0,28	-0,51	0,35	0,76	0,81
2	30	май	1	-0,61	0,45	-0,63	-0,62	0,50	-0,75	0,51	0,68	0,69
		май	2	-0,61	0,42	-0,70	-0,78	0,58	-0,85	0,63	0,82	0,83
		май	3	-0,36	0,03	-0,19	-0,76	0,49	-0,79	0,63	0,73	0,70
		июнь	1	-0,39	-0,11	-0,15	-0,63	0,19	-0,40	0,37	0,54	0,47
3	30	май	1	-0,72	0,57	-0,66	-0,89	0,25	-0,79	0,35	0,66	0,69
		май	2	-0,66	0,35	-0,65	-0,91	0,38	-0,77	0,45	0,88	0,88
		май	3	-0,29	0,16	0,12	-0,75	0,51	-0,55	0,53	0,86	0,87
		июнь	1	-0,20	-0,46	0,19	-0,56	-0,06	-0,12	0,07	0,71	0,72

NDVI – среднее за декаду значение вегетационного индекса; VCI – среднее за декаду значение индекса условий роста растительности; T – средняя декадная температура воздуха; T3 – средняя температура воздуха за 3 декады; P – сумма осадков за декаду; P3 – сумма осадков за 3 декады; D – средний за декаду дефицит влажности воздуха; D3 – средний дефицит влажности воздуха за 3 декады; GTK – значение ГТК за месяц

Мультиколленеарность

Корреляционная матрица, Ростовская область, 2 декада мая

	NDVI	VCI	T	P	D	T3	P3	D3	GTK	Y
NDVI	1									
VCI	0,97	1								
T	-0,59	-0,52	1							
P	0,12	0,23	-0,01	1						
D	-0,65	-0,65	0,75	-0,37	1					
T3	-0,73	-0,63	0,90	0,03	0,70	1				
P3	0,30	0,37	-0,40	0,80	-0,59	-0,38	1			
D3	-0,67	-0,61	0,69	-0,17	0,76	0,87	-0,53	1		
GTK	0,36	0,41	-0,53	0,68	-0,67	-0,51	0,98	-0,64	1	
Y	0,72	0,66	-0,82	0,15	-0,76	-0,90	0,49	-0,84	0,59	1

NDVI – среднее за декаду значение вегетационного индекса; VCI – среднее за декаду значение индекса условий роста растительности; T – средняя декадная температура воздуха; T3 – средняя температура воздуха за 3 декады; P – сумма осадков за декаду; P3 – сумма осадков за 3 декады; D – средний за декаду дефицит влажности воздуха; D3 – средний дефицит влажности воздуха за 3 декады; GTK – значение ГТК за месяц

– Включение в регрессионную модель мультиколлинеарных факторов не совсем корректно.
– В этом случае оценки параметров регрессии не устойчивы.

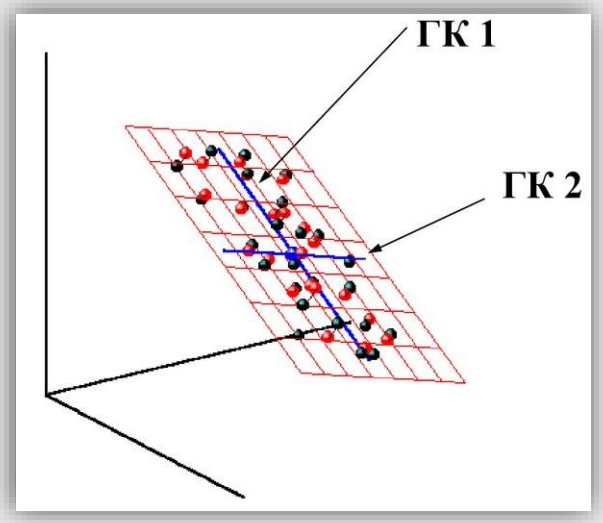
Метод главных компонент

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



- 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

главные компоненты являются линейными комбинациями исходных переменных

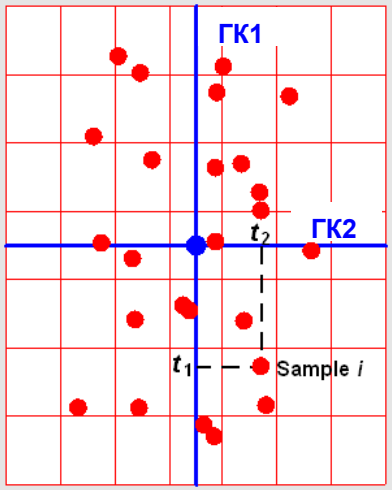
$$T_{iA} = C_1 y_{i1} + C_2 y_{i2} + \cdots + C_j y_{ij} + \cdots + C_p y_{ip}$$
$$T_{(i+1)A} = C_1 y_{(i+1)1} + C_2 y_{(i+1)2} + \cdots + C_j y_{(i+1)j} + \cdots + C_p y_{(i+1)p}$$

. . . . .

$$T_{IA} = C_1 y_{I1} + C_2 y_{I2} + \cdots + C_j y_{ij} + \cdots + C_p y_{ip}$$

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

Проекции исходных переменных на подпространство главных компонент



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

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.

Выбор числа главных компонент:

- 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.
- Отбираются компоненты, чьи собственные значения превышают *1*.

Краснодарский край, 1 декада мая

Для построение компонент использовались следующие параметры: NDVI, VCI, LAI, D3, T3, GTK

Компо-нента	Собствен-ные числа	% общей дисперсии	Кумулят. % общ. дисп.
1	4.33	67.34	67.34
2	1.67	25.95	93.29
3	0.29	4.50	97.80
4	0.11	1.72	99.52
5	0.03	0.40	99.92
6	0.01	0.08	100.00

Краснодарский край, рассчитанная урожайность с 2012 по 2017 гг.

Группа	Относительная ошибка, %			
	1 декада мая	2 декада мая	3 декада мая	1 декада июня
1	5,21	5,59	5,32	8,26
2	7,00	5,27	9,34	11,48
3	6,64	6,43	8,68	10,53

МГК и ПЛС

- Пространство ГК оптимально для внутренней структуры X, но не учитывает структуру Y и связь между X и Y
- ПЛС позволяет учесть связь между X и Y при построении проекционной модели
- ПЛС – пространство создается при участии двух переменных X и Y одновременно, критерием является моделирование той информации в X, которая имеет корреляцию с Y
- ПЛМ-модель специально оптимизирована для регрессионного анализа

Проверка (валидация) модели

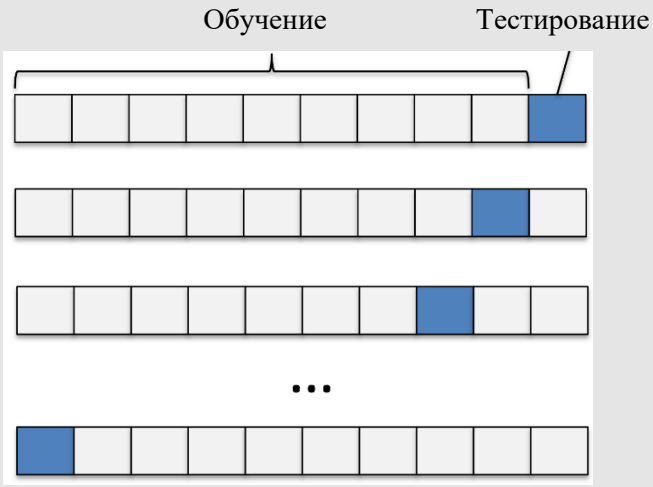
Кросс-валидация .

- Определение размерности модели (числа ГК)
- Оценка предсказательной способности модели

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \overline{y_i})^2}{n}}$$

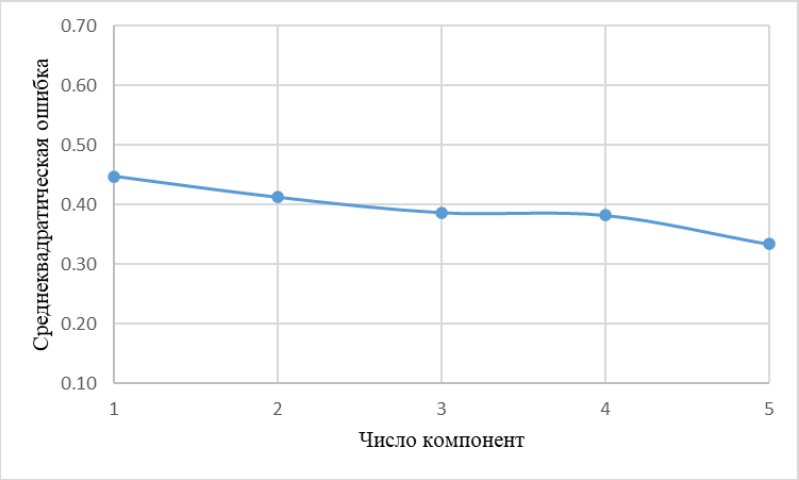
где y_i – фактическое значение урожайности, $\bar{y}_i$ – рассчитанное значение урожайности,

n – число наблюдений.

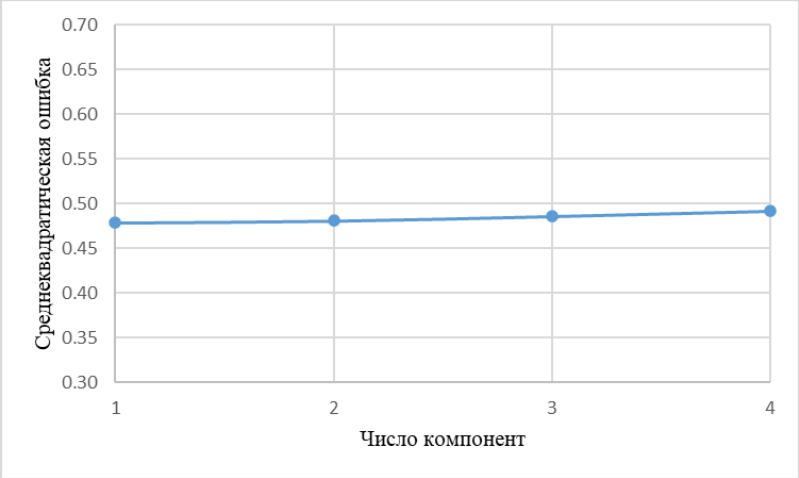


Кросс-валидация

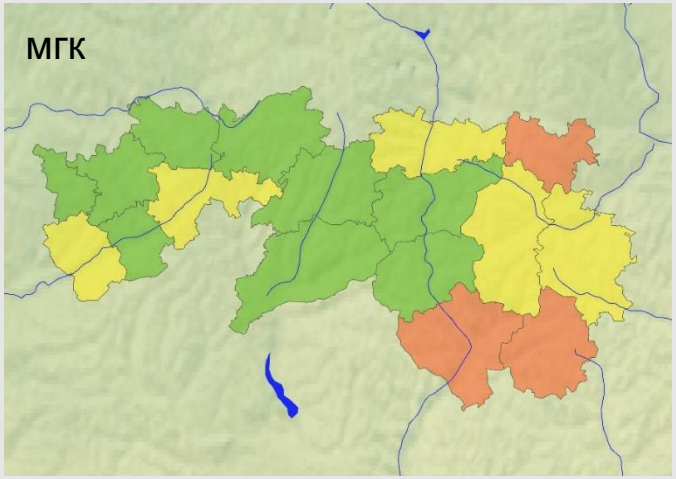
Белгородская область, вторая декада мая
параметры: NDVI, VCI, T, T3, D

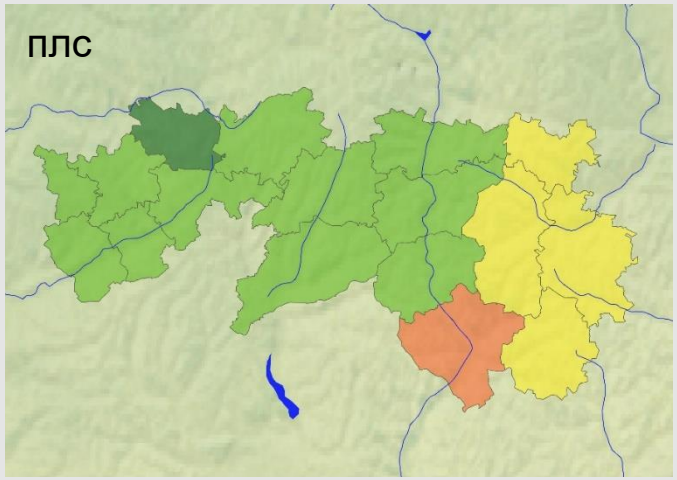


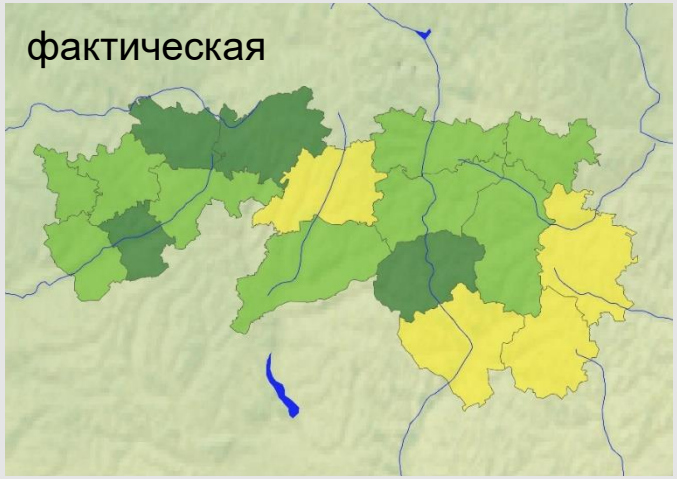
Воронежская область, третья декада мая
параметры: NDVI, VCI, T, D



Белгородская область, 3 декада мая 2017 г.







Урожайность, ц/га

- 39.0 - 44.9
- 44.9 - 50.0
- 50.0 - 55.2
- 55.2 - 60.3

Относительная ошибка, с 2012 по 2021 гг.

Субъект	Группа	1 декада мая		2 декада мая		3 декада мая		1 декада июня	
		Метод							
		МГК	ПЛС	МГК	ПЛС	МГК	ПЛС	МГК	ПЛС
Белгородская	1	7,5	7	7,6	7,4	8,6	8,1	9,2	9,1
Липецкая	1	12,6	11,6	13,3	11,8	12,6	10,4	15,8	14,9
Курская	1	14,7	13,7	12,6	12,6	10,2	10,0	11,6	11,6
Воронежская	1	8,0	8,0	10,0	9,7	11,8	11,1	14,3	11,5
Воронежская	2	11,5	10,5	12,0	11,7	11,5	11,5	14,0	14,0

СПАСИБО ЗА ВНИМАНИЕ