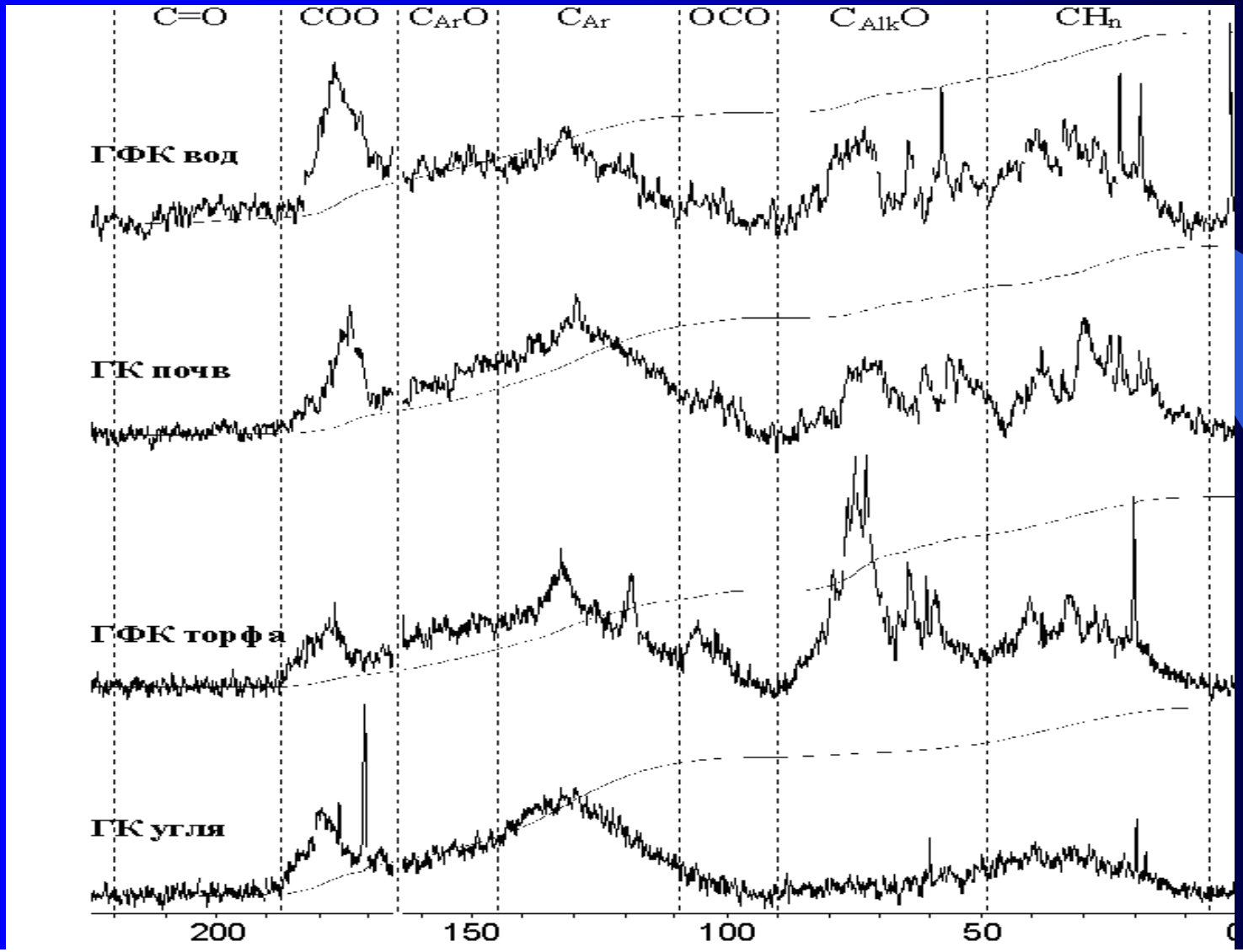
Классификация ГВ по происхождению и фракционному составу методами дискриминантного анализа и «К ближайших соседей»

Константинов А.И.

Количественные характеристики строения гуминовых веществ

- содержание (в масс. %) и атомные соотношения элементов – данные элементного анализа;
- содержание элемента в определённом химическом окружении (в % от общего содержания данного элемента в препарате) – данные ЯМР.

Типичные спектры ЯМР ^{13}C природных ГВ



Характеристики строения ГВ, использованные для классификации

ЯМР: содержание С по спектральным интервалам:	Число интервалов в спектре ЯМР
CH_n, CH_nO, OCO, C_{ar}, $C_{ar}O$, COO, C=O	7
CH_n, CH_3O, CH_2O, CHO, OCO, C_{ar}, $C_{ar}O$, COO, C=O	9
через 12 ppm	19
через 11 ppm	20
через 10 ppm	22
через 9 ppm	25
через 8 ppm	28
через 7 ppm	32
через 5 ppm	44
через 4 ppm	55
через 2 ppm	110

Данные элементного анализа
C, H, N, O, H/C, O/C

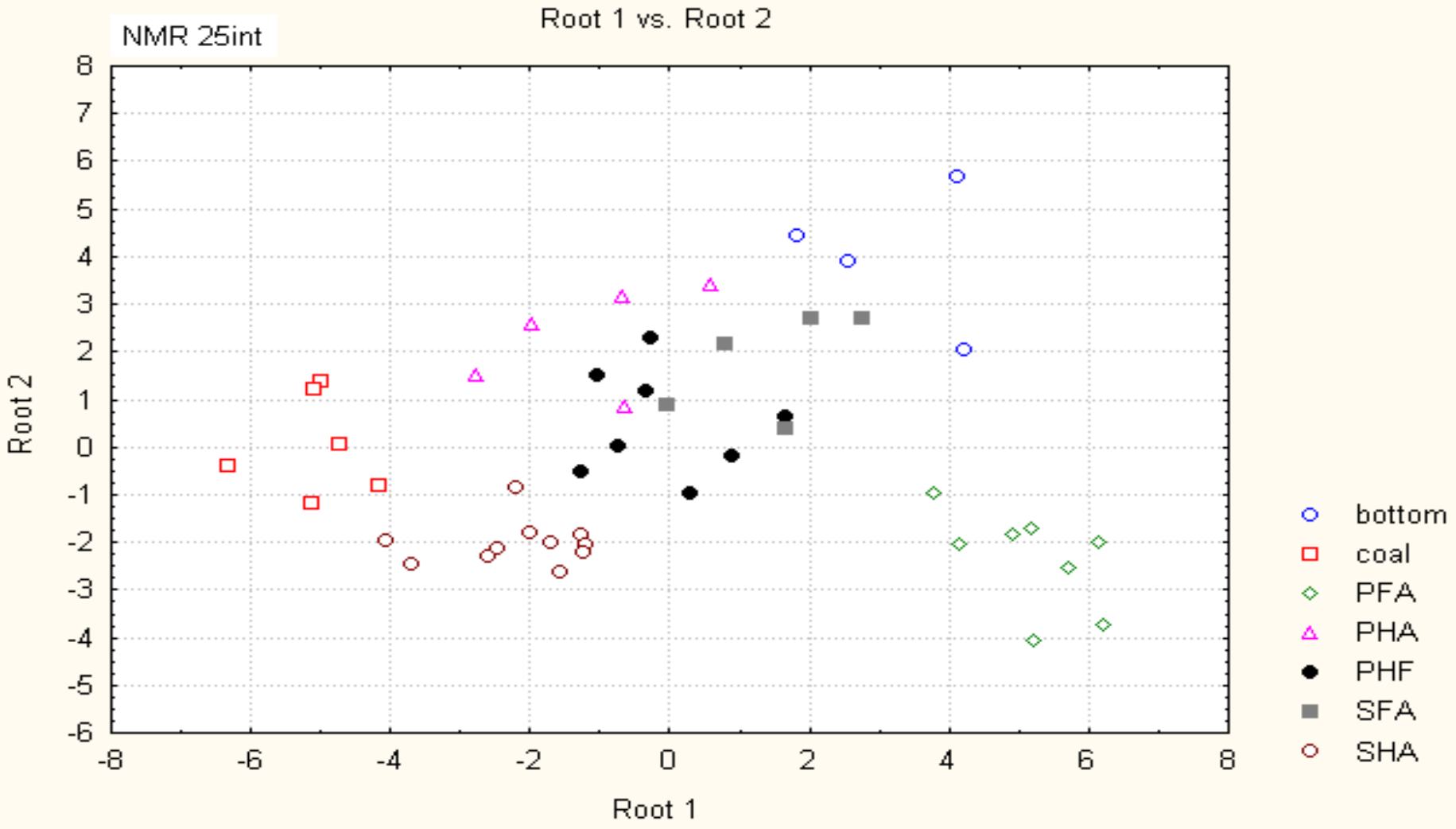
Статистические методы:

- дискриминантный анализ (для построения модели и классификации использовали программу “STATISTICA” компании StatSoft);
- «К ближайших соседей» (с использованием программы “Regression”, автор – А.В. Кудрявцев).

Дискриминантный анализ

1. **Дискриминация** - построение дискриминирующей модели для разделения известной выборки данных на группы.
2. **Классификация** – применение дискриминирующей модели для отнесения новых данных к той или иной группе.

Графическое представление дискриминирующей модели



Дискриминирующие функции

$$Root_i = b_{i0} + b_{i1} \cdot x_1 + b_{i2} \cdot x_2 + \ldots + b_{im} \cdot x_m$$

где:

- x_k – переменные, на основании которых осуществляется разделение групп;
- b_{ik} – подобранные коэффициенты.

Классифицирующие функции

$$S_i = c_i + w_{i1} \cdot x_1 + w_{i2} \cdot x_2 + \ldots + w_{im} \cdot x_m$$

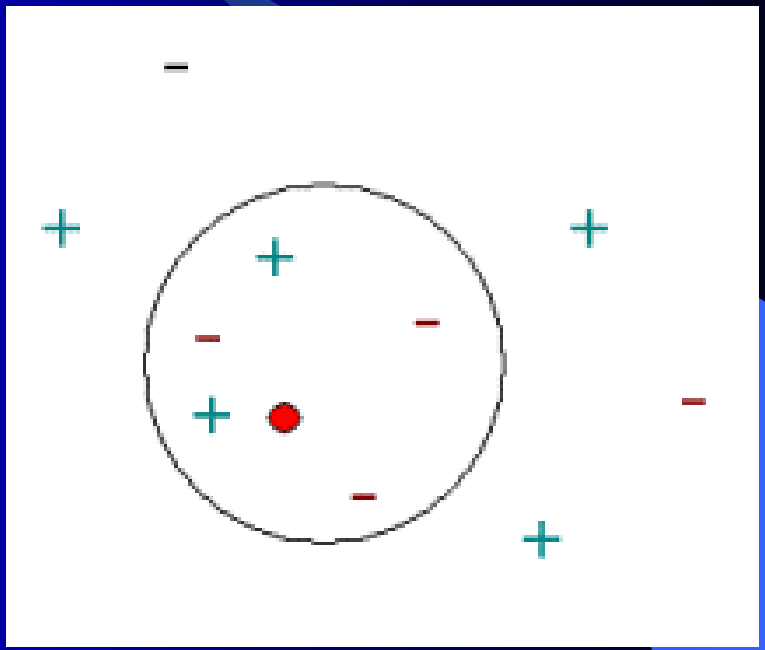
где:

- индекс i обозначает соответствующую группу;
- x_k – переменные;
- w_{ik} - веса классификации для k-й переменной i-й группы;
- c_i – константа для i-й группы;
- S_i – значение классифицирующей функции.

Метод «К ближайших соседей»

Суть метода:

классификация новых данных (точек запроса) на основании их расположения среди известных данных (на рис. красная точка классифицируется как «-», по наибольшему числу соседей, К=5).



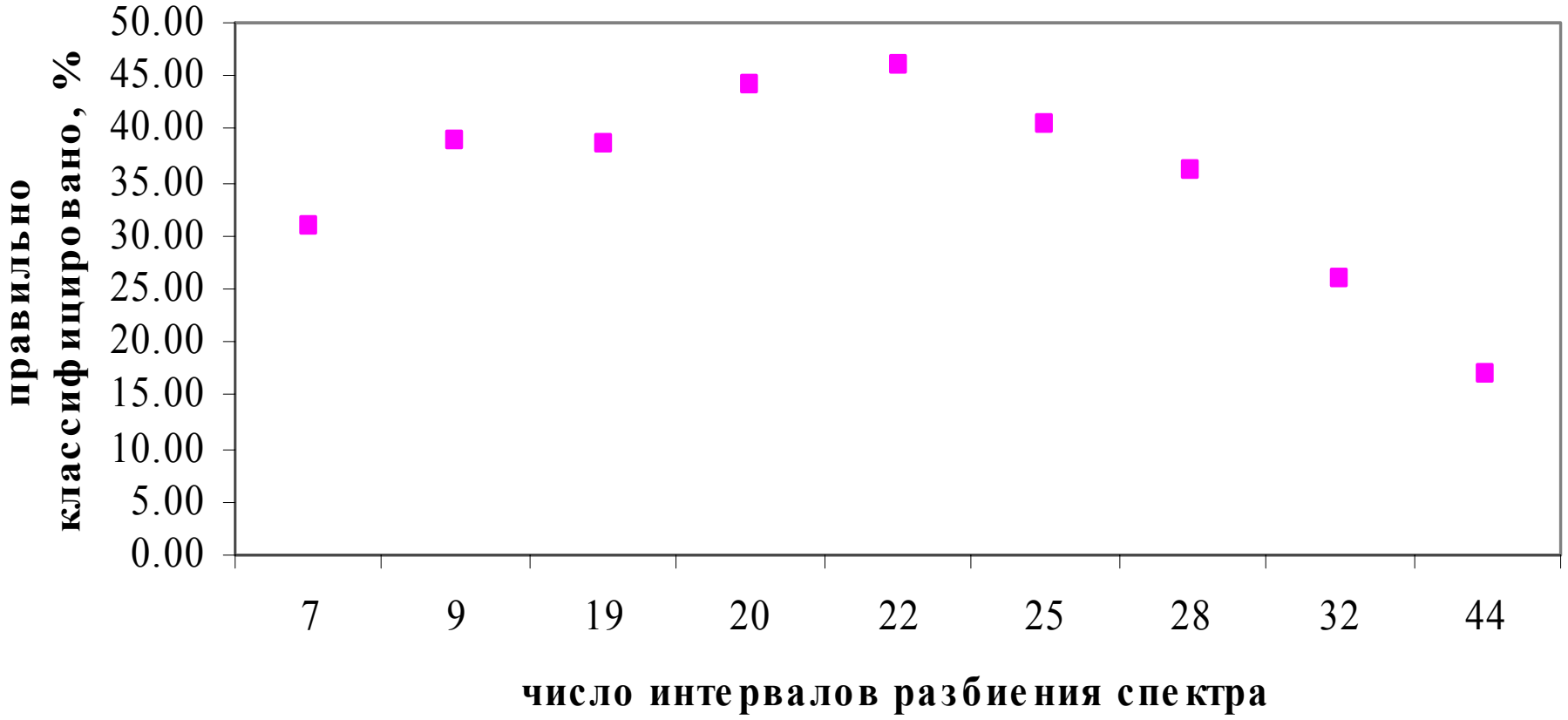
1-й вариант выборки данных

Количество данных в выборке:

	всего	обучающая	контрольная
aquatic	3	2	1
bottom	5	2	3
coal	9	5	4
PFA	12	6	6
PHA	8	4	4
PHF	13	6	7
SFA	8	4	4
SHA	16	8	8
SHF	4	2	2
Summa	**78**	**39**	**39**

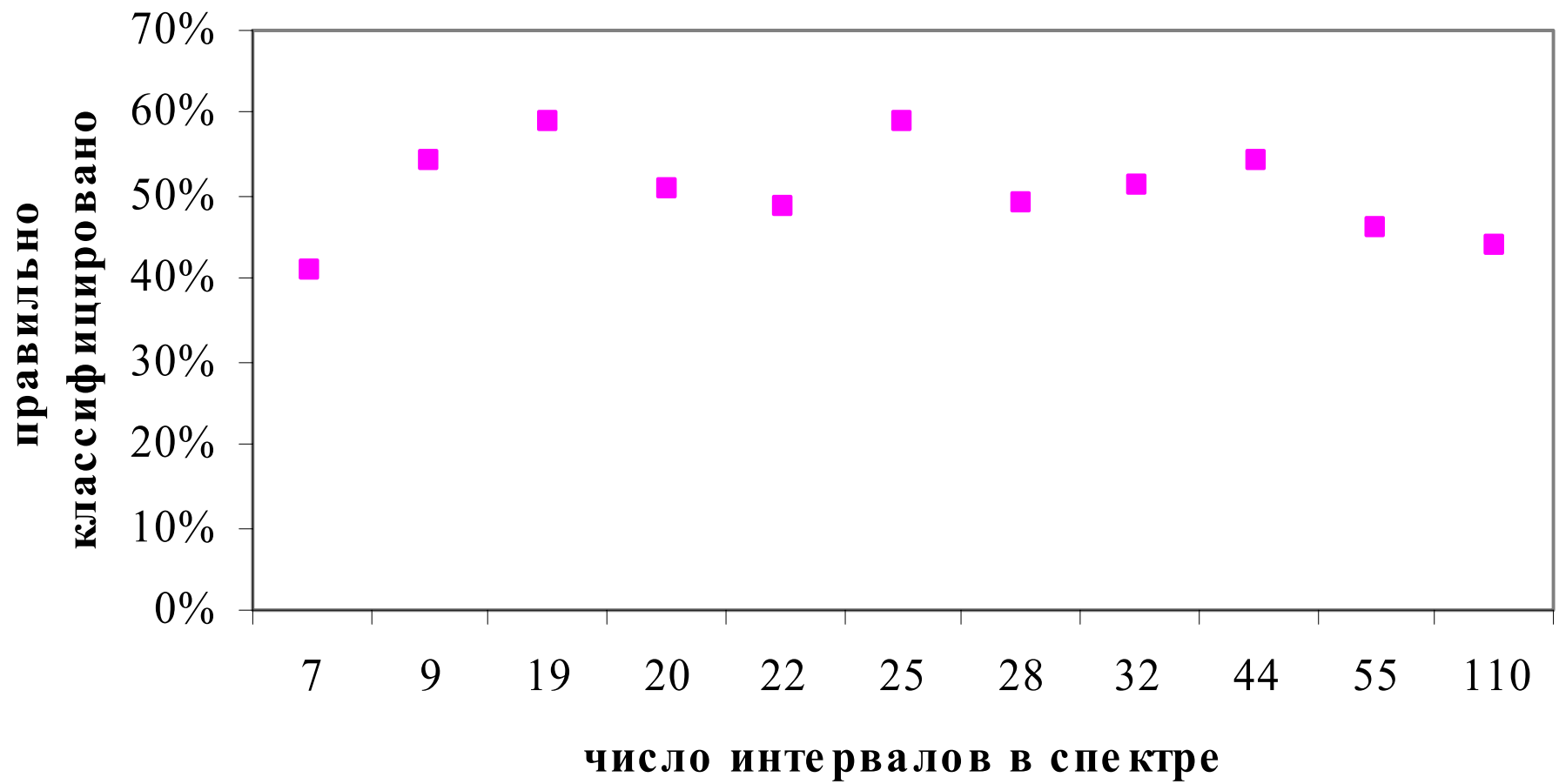
Результаты классификации по 1-му варианту

Методом дискриминантного анализа



Результаты классификации по 1-му варианту

Методом «*K* ближайших соседей»

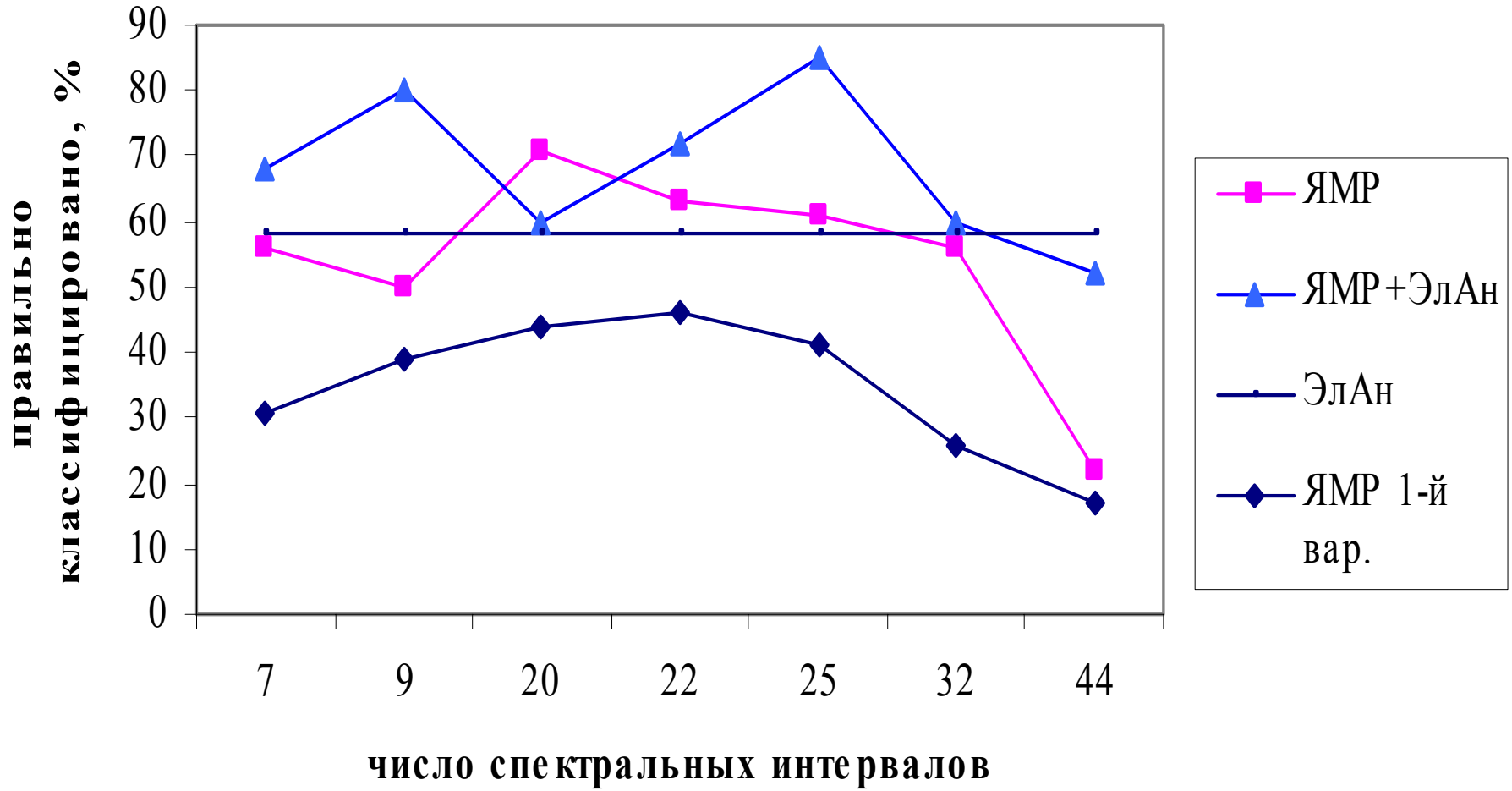


2-й вариант выборки данных

ЯМР			
	всего	обучающая	контрольная
bottom	5	4	1
coal	9	6	3
PFA	12	8	4
PHA	8	5	3
PHF	13	8	5
SFA	8	5	3
SHA	20	12	8
Summa	**75**	**48**	**27**
ЭлАн+ЯМР			
	всего	обучающая	контрольная
bottom	5	4	1
coal	9	6	3
PFA	12	8	4
PHA	8	5	3
PHF	13	8	5
SFA	6	3	3
SHA	16	10	6
Summa	**69**	**44**	**25**

Результаты классификации по 2-му варианту

Методом дискриминантного анализа

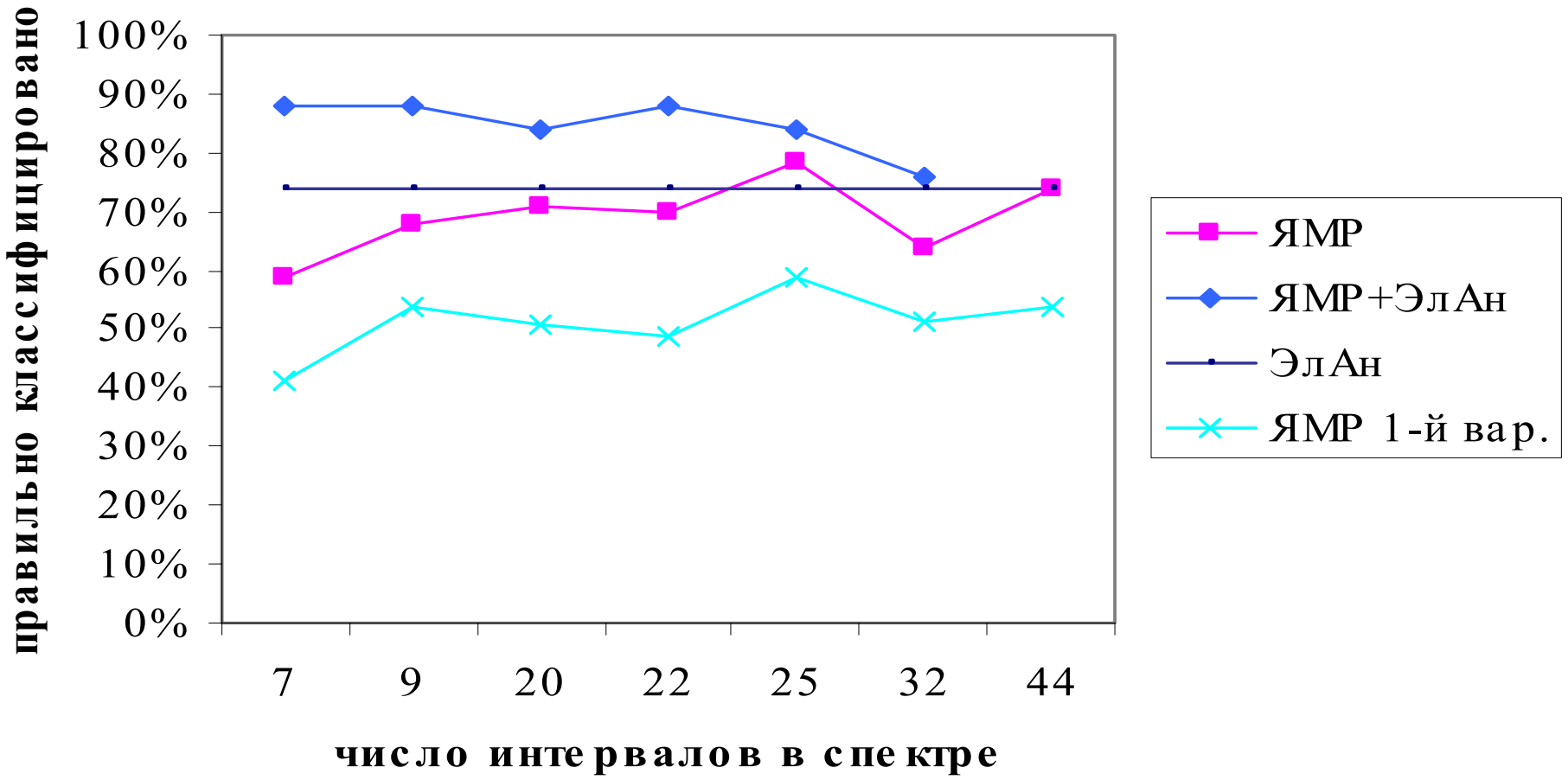


Характеристики классификации препаратов ГВ методом дискриминантного анализа

Набор данных	Правильно классифицировано	Дескрипторы
Эл.Ан.	58%	N, H/C, O/C, H, C
ЯМР ^{13}C, 7 интервалов	56%	C_{AR}, CH_nO, COO, OCO, $C_{AR}O$
ЯМР ^{13}C, 20 интервалов	71%	132-143, 44-55, 66-77, 165-176, 55-66, 143-154, 154-165, 33-44, 110-121, 187-198, 22-33, 198-209, 121-132 ppm
Эл.Ан. + ЯМР ^{13}C, 9 интервалов	80%	H/C, N, C_{AR}, O/C, CH_2O, C, O, CHO, COO
Эл.Ан. + ЯМР ^{13}C, 25 интервалов	85%	H/C, N; 60-69, 168-177 ppm; O/C; 105-114, 195-204, 186-195, 33-42, 78-87, 69-78, 141-150, 87-96, 114-123, 42-51, 123-132, 51-60 ppm

Результаты классификации по 2-му варианту

Методом «К ближайших соседей»



Характеристики классификации препаратов ГВ методом «К ближайших соседей»

Набор данных	Правильно классифицировано, %	Дескрипторы
Эл.Ан.	74	H, N, O, H/C, O/C
ЯМР ^{13}C, 9 интервалов	68	CH_3O, CHO, C_{AR}, $C_{AR}O$
ЯМР ^{13}C, 25 интервалов	79	24-33, 60-78, 150-159, 168-177 ppm
Эл.Ан. + ЯМР ^{13}C, 9 интервалов	88	C_{AR}, N, H/C
Эл.Ан. + ЯМР ^{13}C, 25 интервалов	84	N, H/C; 123-132 ppm

Выводы:

- классификация ГВ по происхождению и фракционному составу методами ЛДА и КБС даёт наилучший результат при совместном использовании в качестве характеристик состава ГВ данных элементного анализа и ЯМР ^{13}C при разделении спектра на 9 интервалов, соответствующих различным структурным фрагментам ГВ, и при разделении спектра на интервалы через 9-11 м.д.;
- в дальнейшем предполагается расширить выборку препаратов ГВ и привлечь другие дескрипторы состава для более детальной и более точной классификации.