Област на приложимост на модел

Защо се нуждаем от област на приложимост?

Какво е област на приложимост?

Методи за определяне на покритието на обучителната извадка

Идентификация на областта на предсказване

Практическа употреба

QSAR моделиране

QSAR/QSPR моделите се прилагат в процесите на вземане на решения във фармацевтичната и химичната индустрия.

Като резултат се получава:

- по-голяма ефективност на финансовите разходи

- съкращаване на времето за разработка

- откриванене на алтернативи за многобройните тестове с животни

- методи за разработка щадящи околната среда и природните ресурси

QSAR моделиране

Основен проблем е оценката на качеството на моделите и превенция от неправилна употреба на тези методи.

Приемането на резултат = предсказване от областта на приложимост

Елементи на качествено предсказване

- определяне дали моделът е подходящ за даденото съединение

- оценяване на неопределеността на резултатите от модела

QSAR моделиране

Световната организация за икономическо сътрудничество и развитие (OECD) насърчава употребата на QSAR/QSPR модели.

За целта след 2000 г. се въвеждат официални изисквания за качеството на QSAR моделите.

В продължение на повече от 40 години за разработени и публикувани хиляди QSAR/QSPR модели.

Има голямо затруднение в преценката, кои модели могат да се използват практически.

QSAR моделиране

В миналото QSAR изследователите се фокусираха върху анализа на данните и получаването на самите модели

Областта на приложимост не се е изследвала.

Приемането на QSAR модела се оставя на експертната оценка на потребителя на този модел.

QSAR моделиране

През 2004 г. OECD въвежда 5 основни принципа за валидиране на QSAR модел описани в официален документ: **OECD Principles for (Q)SAR Validation**

- ясно дефинирано целево свойство

- недвусмислен алгоритъм за моделиране

- дефинирана област на приложимост

- подходящи статистически характеристики доказващи качеството на приближението, стабилността и предсказ-вателната способност на модела

- “механистична” интерпретация на модела **(ако е възможно)**

Стандартни целеви свойства

Физико-химични

Melting Point

Boiling Point

Vapour Pressure

K octanol/water

K organic C/water*

Water Solubility

Стандартни целеви свойства

Околна среда

Biodegradation

Hydrolysis

Atmospheric Oxidation

Bioaccumulation*

Екологични ефекти

Acute Fish Toxicity

Acute Daphnid Toxicity

Alga Toxicity

Long-term Aquatic Toxicity

Terrestrial Effects

Стандартни целеви свойства

Ефекти касаещи човешкото здраве

Acute Oral Toxicity

Acute Inhalation Toxicity

Acute Dermal Toxicity

Skin Irritation /Corrosion*

Eye Irritation/Corrosion *

Skin Sensitisation *

Repeated Dose

Genotoxicity (in vitro)

Genotoxicity (in vitro, non bacterial)

Genotoxicity (in vivo)

Reproductive Toxicity

Developmental Toxicity

Carcinogenicity*

Organ Toxicity (hepatotoxicity, cardiotoxicity, nephrotoxicity, ...)

Необходимост от “област на приложимост”

Необходимостта от **Област на Приложимост** (ОП) се обуславя от факта, че QSAR са редуциращи модели т.е. моделите се асоциират с ограничения по отношение на:

- химичните структури

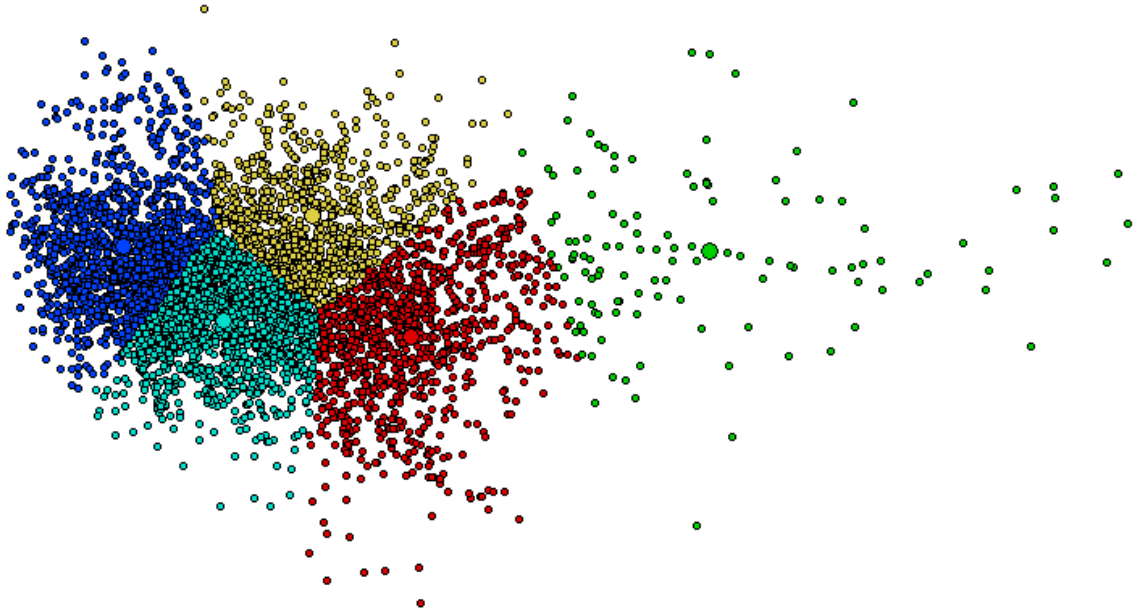
- физико-химичните свойства

- и механизмите на взаимодействие, за които могат да се направят достоверни предсказвания

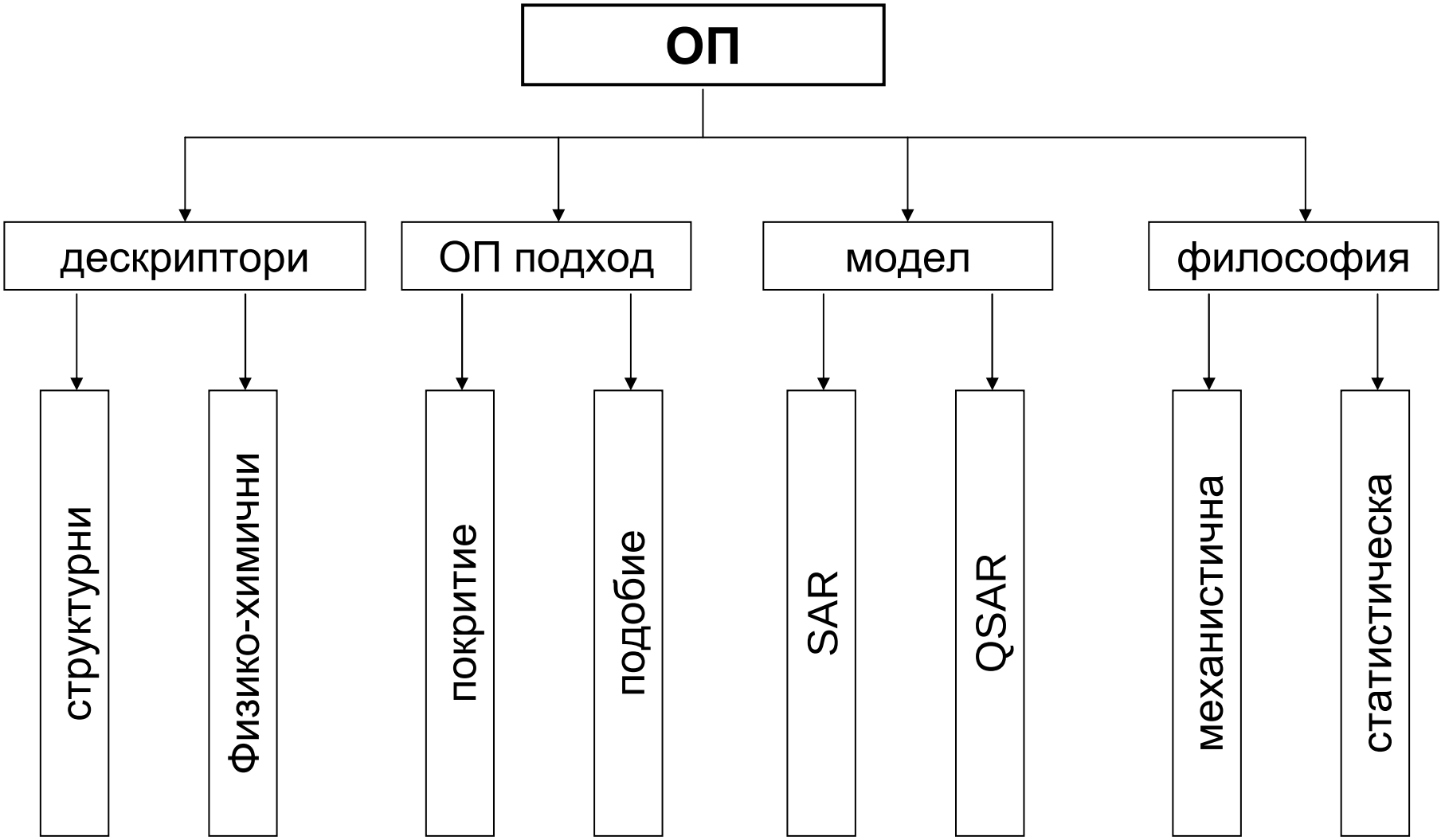
Област на приложимост (ОП)

Дефиниция:

Областта на приложимост за даден модел е част от химичното пространство, където моделът предсказва целевото свойство с определно ниво на достоверност.



Аспекти на ОП



Механистична интерпретация

Терминът **механистичен** първоначално е бил свързан с QSPR моделиране на токсичност.
Фразата “различен **механизъм** на токсично взаимодействие” се използвала за обозначение на бегълци в дадено множество.

Понастоящем броят на потенциално използваемите дескриптори е много голям.

Терминът “механистичен” се отнася за модели, при които използваните дескриптори може да се интерпретират в смисъл на информацията, която кодират за структурата и физико-химичните свойства.

Механистична интерпретация

Моделите с механистична интерпретация са много малко.

Примери за “интерпретируеми” дескриптри:

- коефициент на разпределние на $\mathbf{logK_{o/w}}$ между фазите октанол и вода.

- енергията на най-ниската незаета молекулна орбитала $\mathbf{E_{LUMO}}$

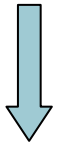
Механистична интерпретация

Принципно когато моделът е с механистична интерпрета-ция може да се дефинира ОП с много по-малки извадки от съединения в сравнение със статистическите подходи.

Налице е хипотеза за определен механизъм на взаимодействие и се подбират специфични съединения, с които да се тества тази хипотеза. По-този начин областта (ограниченията) за употребата на модела може да се получи с по-малко съединения.

Статистически подход

При разработването на чисто статистически модели няма никакви предположения за причината на изследвания ефект / целево свойство (*анг. end point*).



Необходимо е да се тестват голям брой съединения за да се определят вариациите на отделните дескриптори преди да се използва статистика за дефиниране на ОП.

Пример за модел с механистична интерпретация:

“аква-токсичност” (aquatic toxicity)

Повечето индустриални химикали имат токсично десйтвие върху водните живи организми основан на механизма на наркозата (не-валентни и обратими взаимодесйтвия)

Механистични QSPR модели

Октанола се приема като заместник (представител) на липидите в изследваните живи организми.

Коефициентът на разпределение **$\log K_{ow}$** се използва като основен дескриптор за описание на взаимодействието на даден химикал с клетъчната мембрана:

$$EC_{50} = f(\log K_{OW})$$

регресионен метод

експертна система

Механистични QSPR модели

При разширяване на множеството на моделираните съединения (обхвата на модела) трябва да се изследват съединения, които показват остра токсичност - много по-голяма от предсказаните стойности чрез моделите базирани на механизма на “наркозата”.

Включват се допълнителни дескриптори:

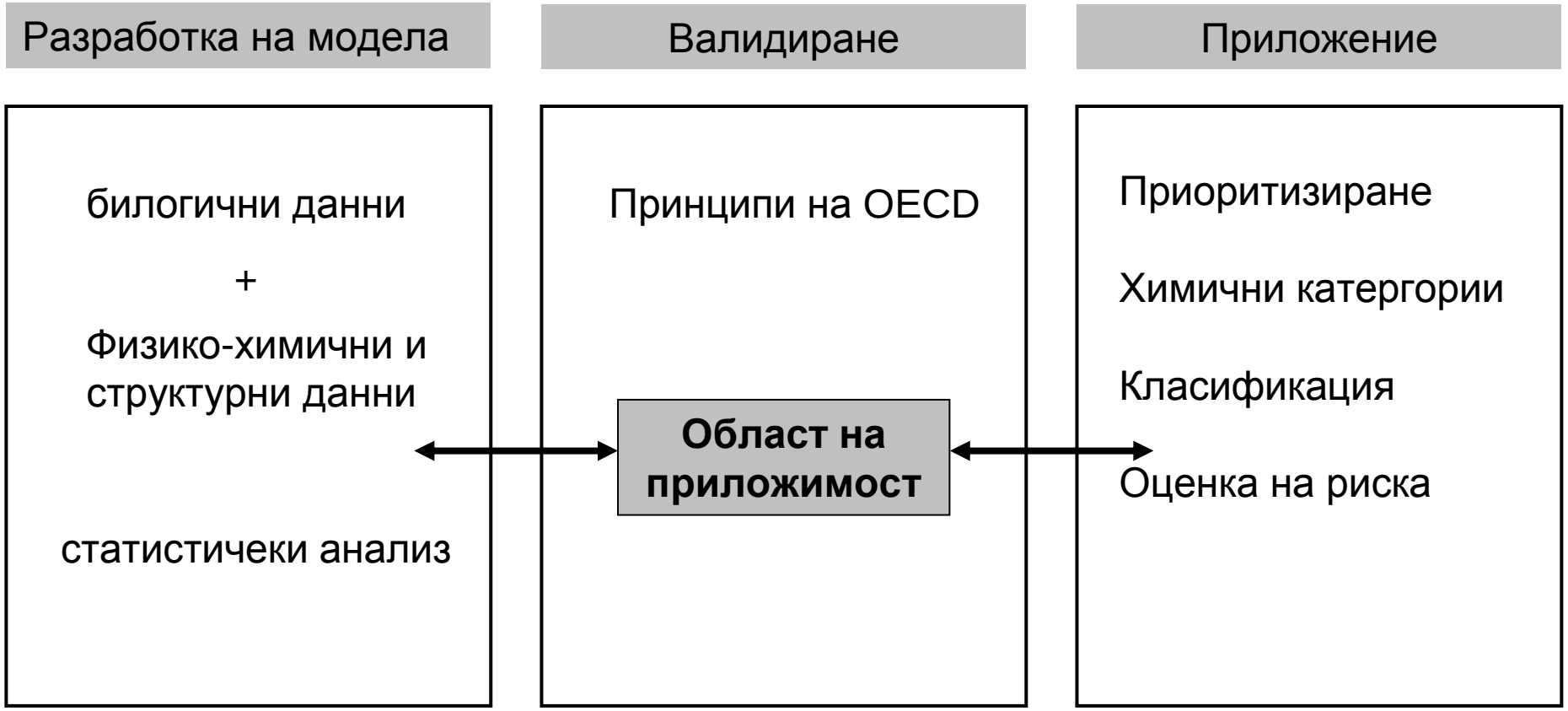
$$EC_{50} = f(\log K_{OW}, d_1, d_2, ..., d_k)$$

Подобрение на статистическите показатели и разширене на ОП

Намаляване на механистичната интерпретация

Значимост на ОП

ОП играе важна роля в трите основни етапа на QSAR моделирането

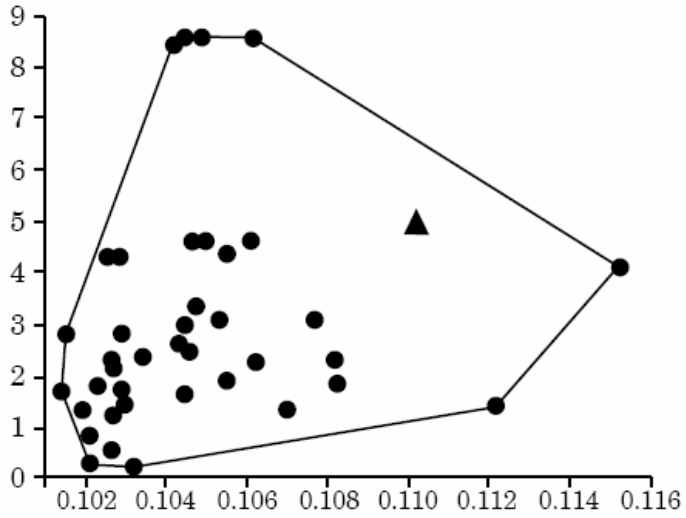


Определяне на ОП

Обучителната извадка, от която е изчислен QSAR моделът е основата за оценяването на ОП.

Данните от обучителната извадка се проектират в многомерното пространство дефинирано от дескрипторите на модела.

В пространството на дескрипторите се оформят региони запълнени с данни (обекти) и празни региони.

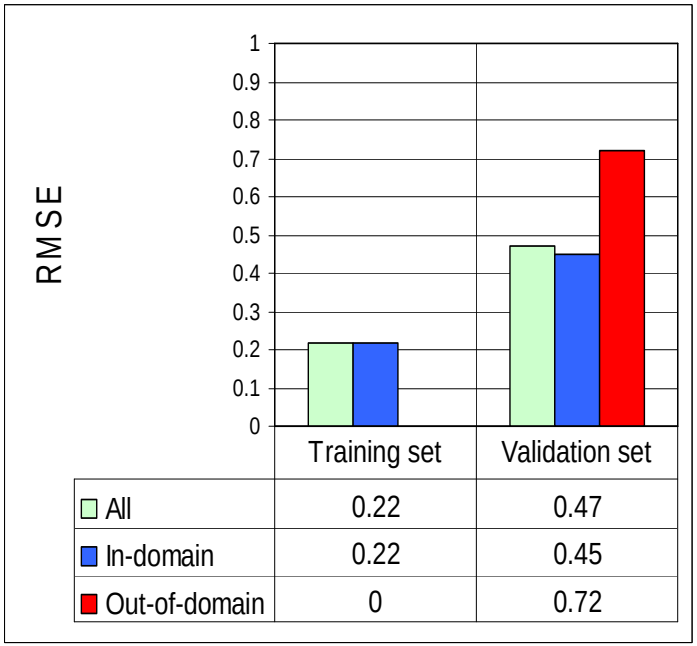


Запълнените региони дефинират ОП на модела – това кореспондира с принципа, че интерполацията е по-надежна от екстраполацията

Интерполация с/у екстраполация

"Интерполиращата" предсказвателна точност се дефинира в рамките на обучителната извадка.

"Екстраполиращата" предскавателна точност се определя чрез съединения извън обучителната извадка, която определя ОП.



Средната грешка за съединенията извъм ОП е около 2 пъти по-голяма в сравнение със съединенията от ОП.

Това наблюдение е вярно само в термините на "средно статистическо" наблюдение. Има индивидулани случаи с ниска грешка извън ОП и висока грешка в ОП.

Други изисквания за ОП

В досегашните разглеждания ОП се основава **само** на областите от химичното пространство покрити от обучителната извадка.

Въпрос:
Ако две различни целеви свойства се моделират с една и съща обучутелна извадка, областите на приложимост на двата модела ще бъдат ли еднакви?

При определяне на ОП има нужда да се разглежда и самият модел.

Други изисквания за ОП

Прилагането на модела за предсказване в интерполационна област не подсигурява високо ниво на точност на модела.

- единствената връзка към модела са стойностите на дескрипторите

Необходима е оценка за предсказващата способност за да се приеме резултатът от моделирането като коректен.

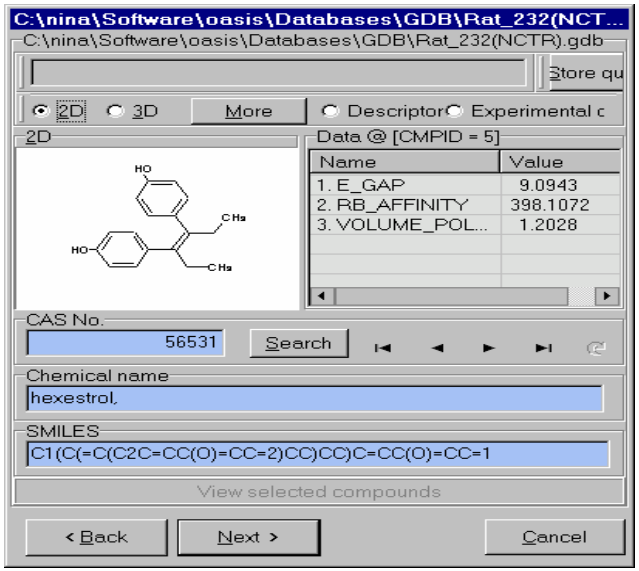
Грешката на моделирането зависи от:

- вариацията на експерименталните стойности

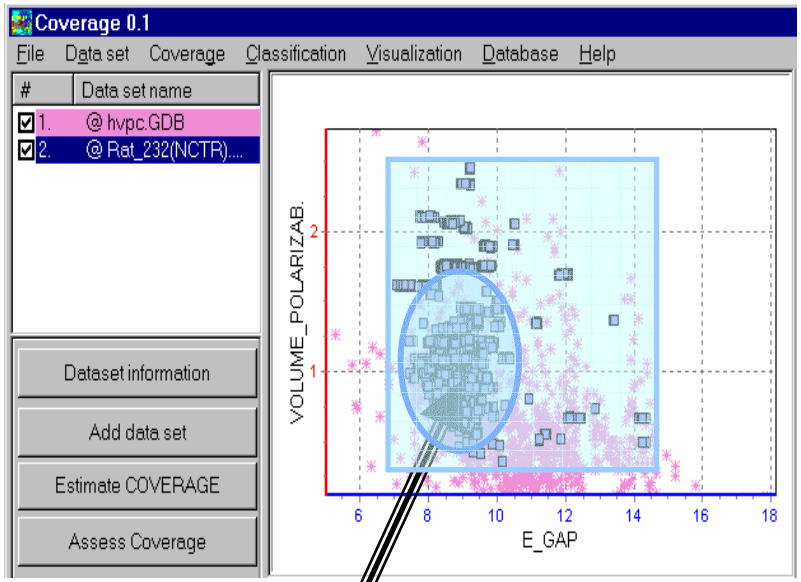
- неопределонността на параметрите

ОП на QSAR модел

Обучителна извадка



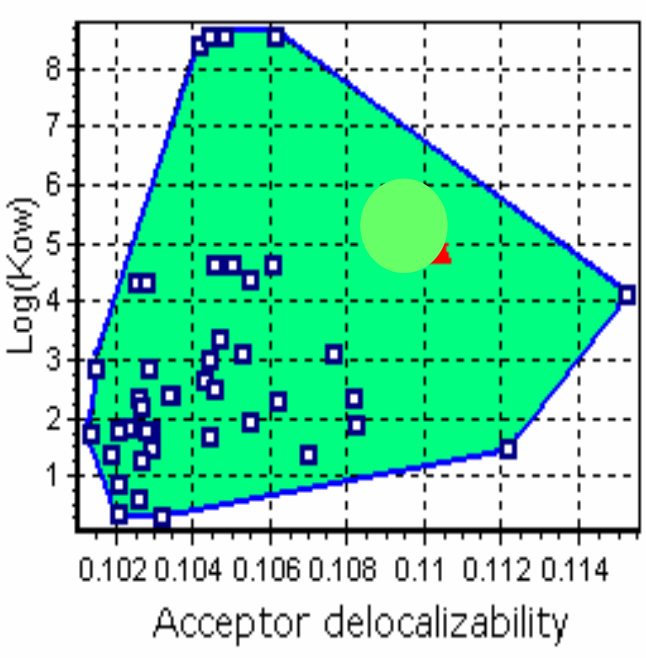
Многомерно пространство от дескрипторите



ОП = запълнените региони в пространството ?

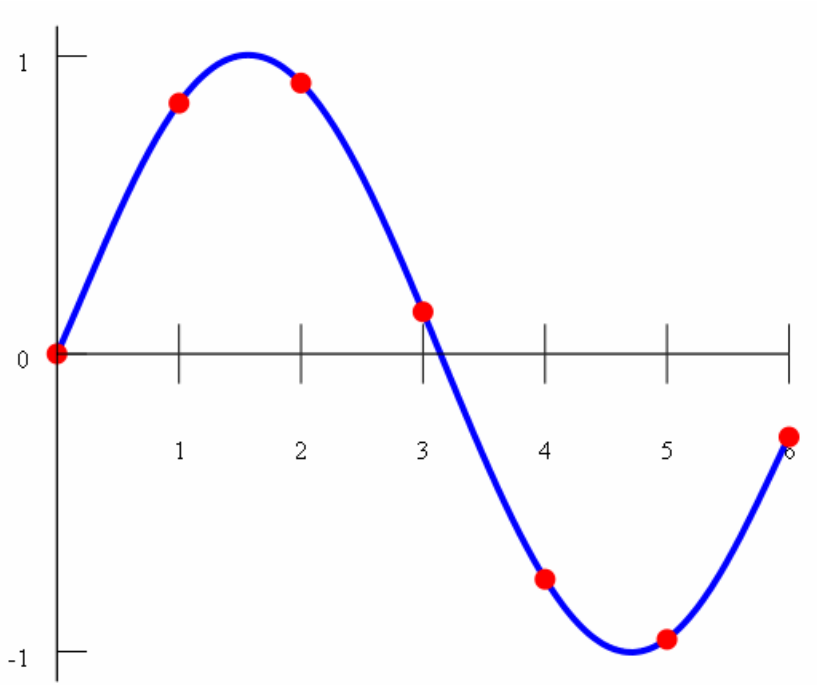
ОП на QSAR модел

- Повечето QSAR модели не са LFER модели
- По същество моделите са статистически с променливо ниво на механистична интерпретация определна след създаването на модела (posteriori)
- Статистическите модели се ограничават до интерполационния регион дефиниран от обучаващата извадка



Интерполация

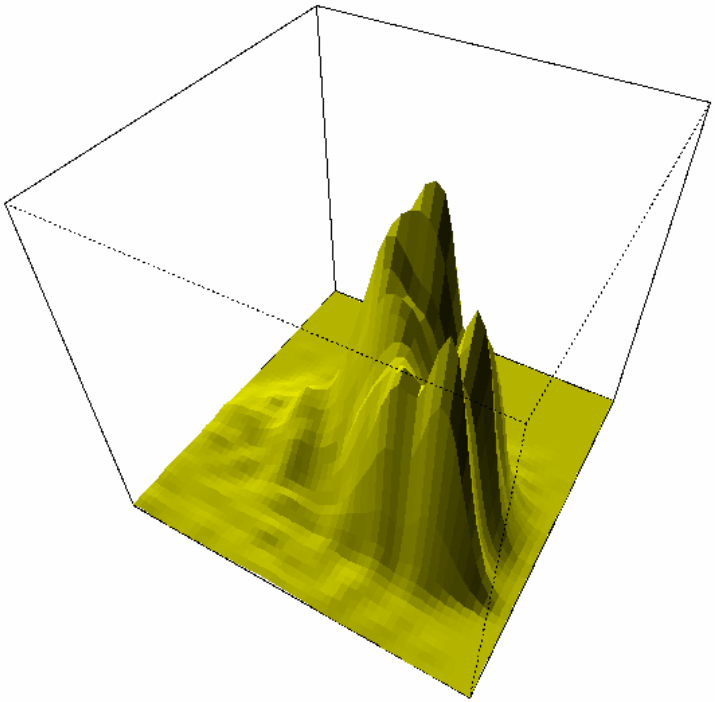
В математиката интерполация се нарича определянето (предсказването) на стойностите на нови точки в рамките на дискретно множество точки с известни стойности

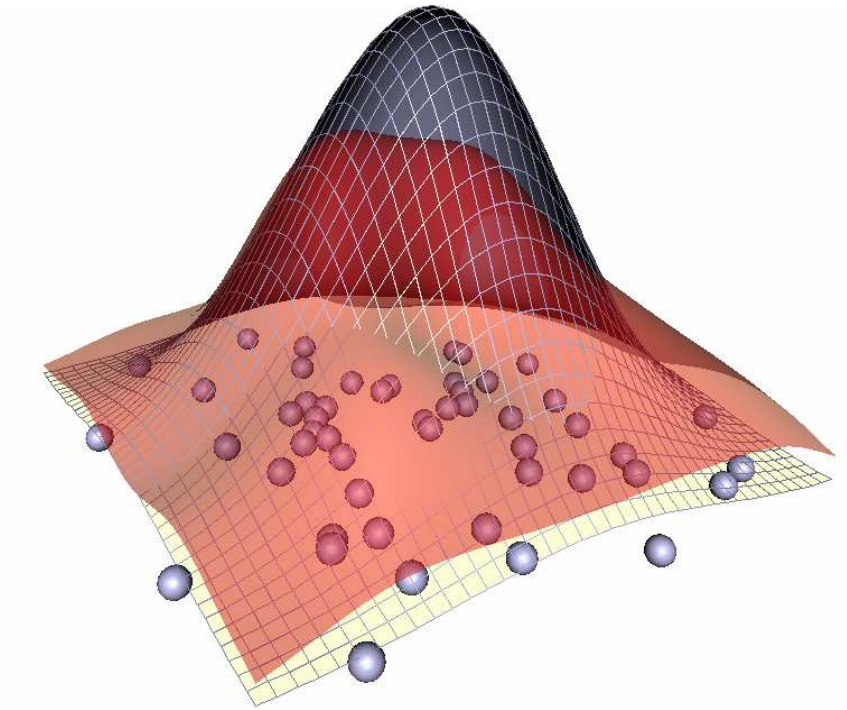


Интерполация

Интерполационния регион в едномерното пространство е **обикновен интервал** $[d_{min},d_{max}]$

В многомерносто пространство интерполационните региони са по-сложни.

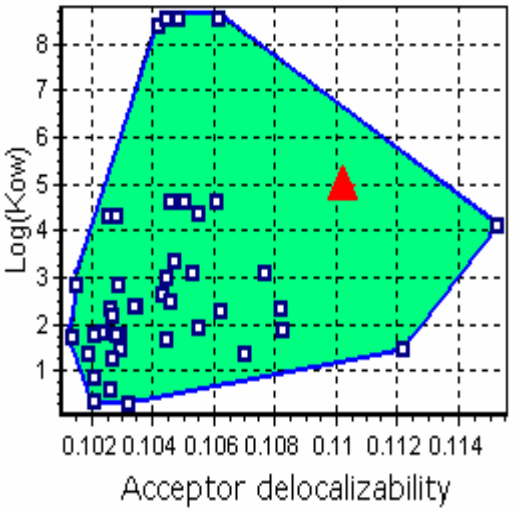




Интерполационен регион

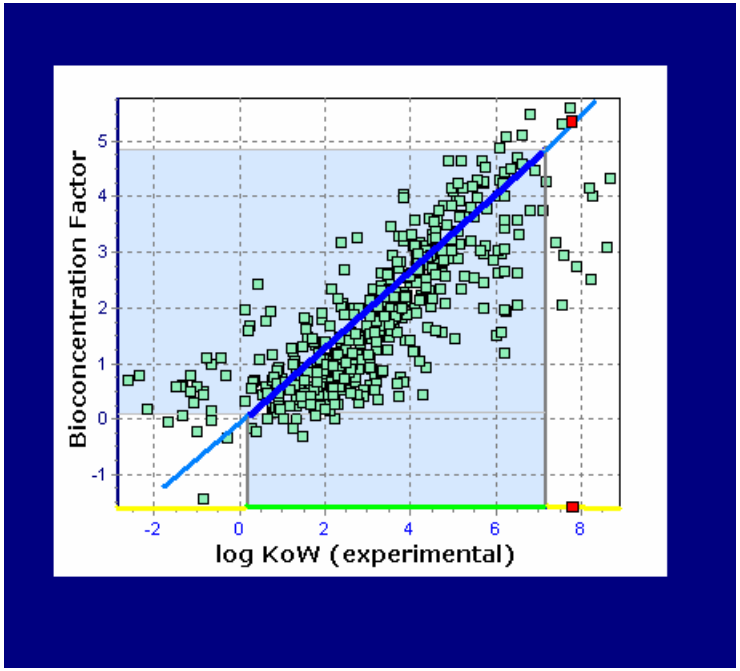
В реалноста данните са разпръснати и нехомогенни:

- Методите базирани на адитивни схеми от по-висок ред (Group contribution methods) имат особенно се влошени регионни в многомерното пространство
- Данните от обучителната извадка не могат да се подберат да следват точно експерименталния дизайн защото начинът на оценяване на свойствата е ретроспективен.



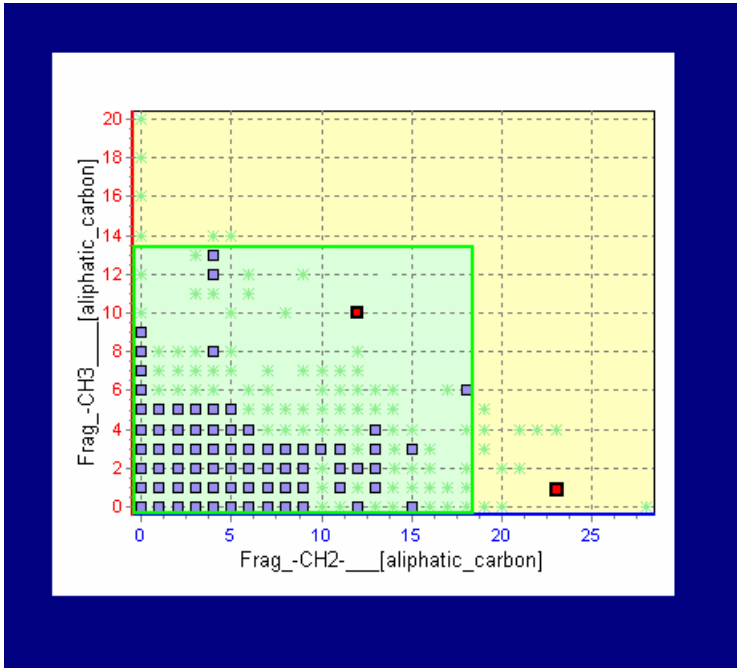
Възможно е да съществуват празни региони в интерполационната област. Взаимовръзките между празните регионни може да различават от изведения модел.

Интерполация с/у Екстраполация



Легенда:

BCF обучаваща извадка



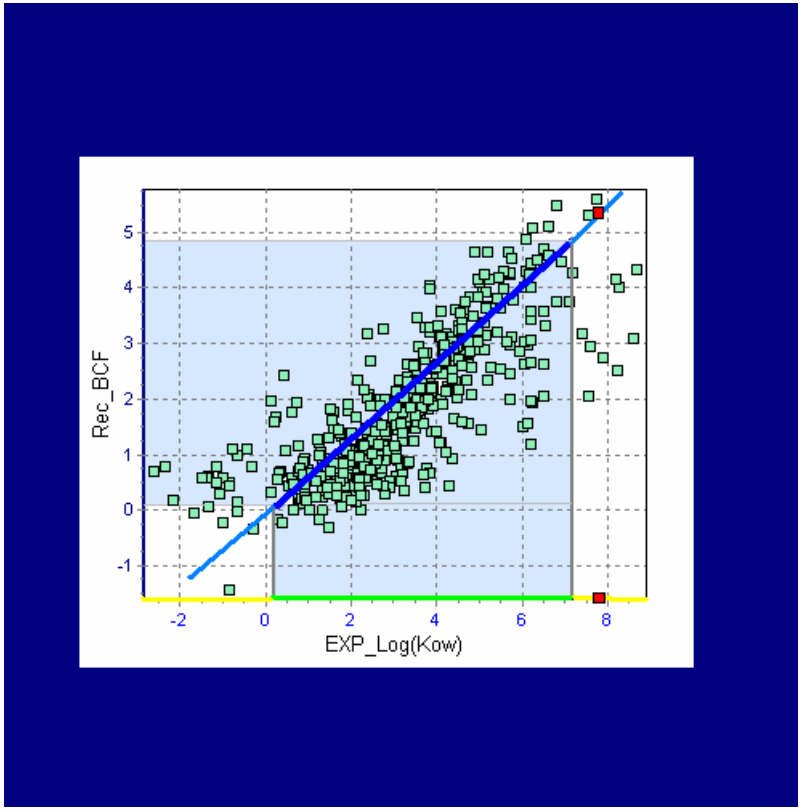
Легенда:

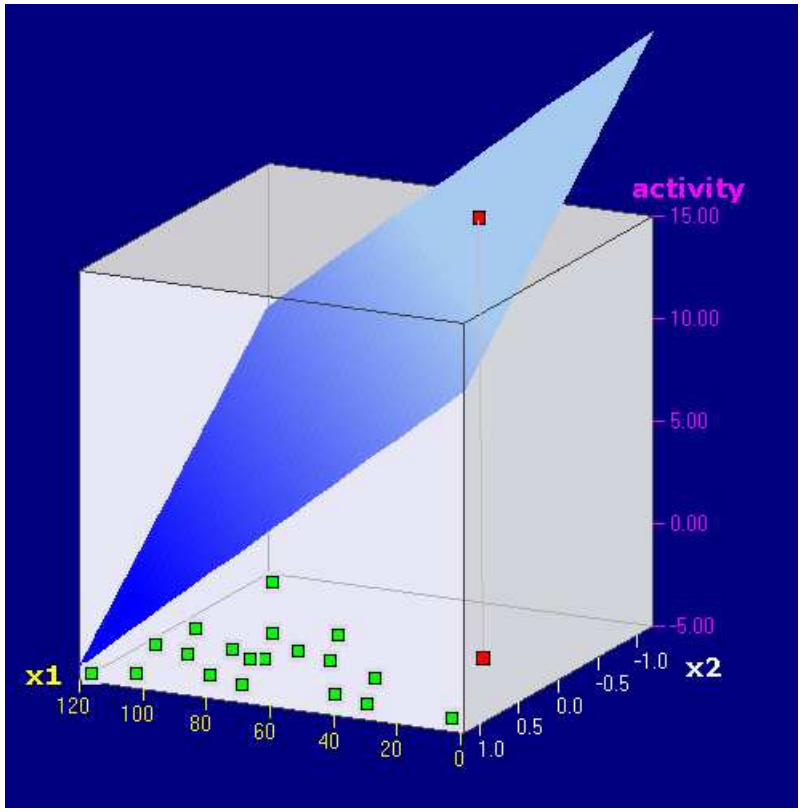
SRC KOWWIN обуч. изв.

1D: интервал определя интерполационната област

2D: Съществуват ли празни региони в интерп. област ?

Интерполация с/у Екстраполация





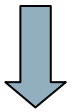
Предсказаните стойности в интерполационния регион не се отклоняват много от експерименталните стойности

При линеен модел (2D), предсказаните стойности е възможно да се отклонят драстично от експеримента-лните дори в интерполацнинния регион

ОП чрез интерполационен регион

Определянето на многомерна област в общ вид чрез интевали не е достатъчно адекватно описание интерполационнния регион:

$$[d^1_{\min}, d^1_{\max}] \times [d^2_{\min}, d^2_{\max}] \times \ldots \times [d^n_{\min}, d^n_{\max}]$$

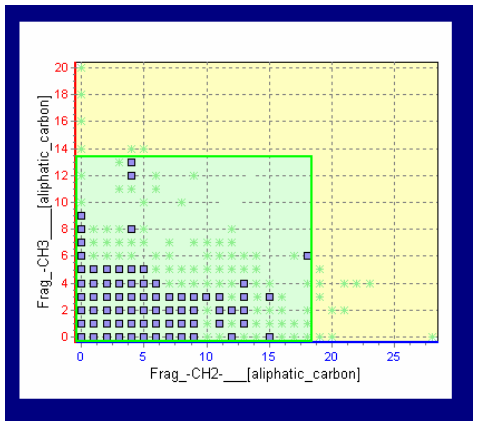


Извод:

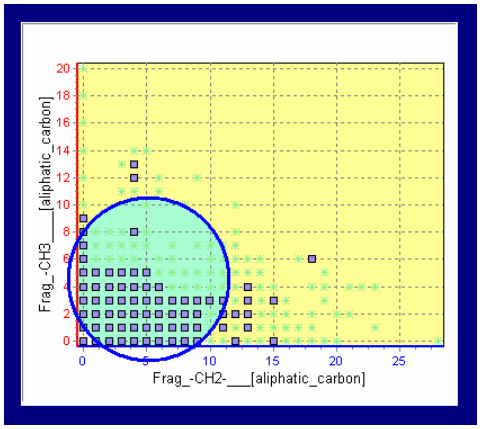
Необходими са по-прецизни методи за оценка на интерполационния регион на даден модел!

Методи за оценка на интерполационен регион

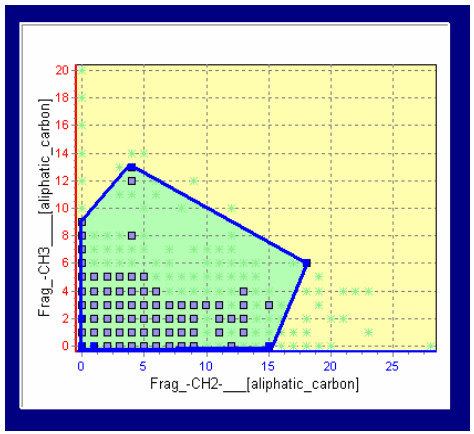
- **Интервали на дескрипторите**



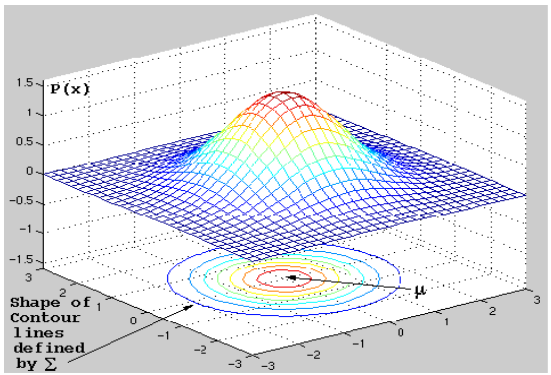
- **Метод на разстоянията**



- **Геометричен метод**



- **Плътност на разпределение**



Метод на интервалите

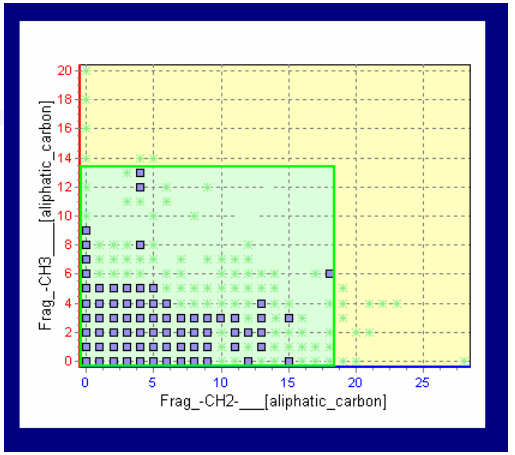
- този метод разглежда интервалите на отделните дескриптори;

-дефинира се n-размерна правоъгълна област със страни успоредни на координатните оси.

- областа се дефинира като декартовото произведение на интервалите за отделните дескриптори:

$$[x^1_{\min}, x^1_{\max}] \times [x^2_{\min}, x^2_{\max}] \times \ldots \times [x^n_{\min}, x^n_{\max}]$$

Метод на интервалите



- Много лесен за реализация
- Работи за модели с висока размерност
  - Единственото практично приложение за груповите адитивни методи
- Не може да се засечат “дупките” в интерполационния регион
- Предполага се хомогенно разпределение на данните
- Не може да се направят корекции за корелациите между дескрипторите

Метод на интервалите - PCA

Вместо ориниалните дескриптори се използват главните компоненти.

$$\{x_i\} \xrightarrow{PCA} \{u_i\}$$

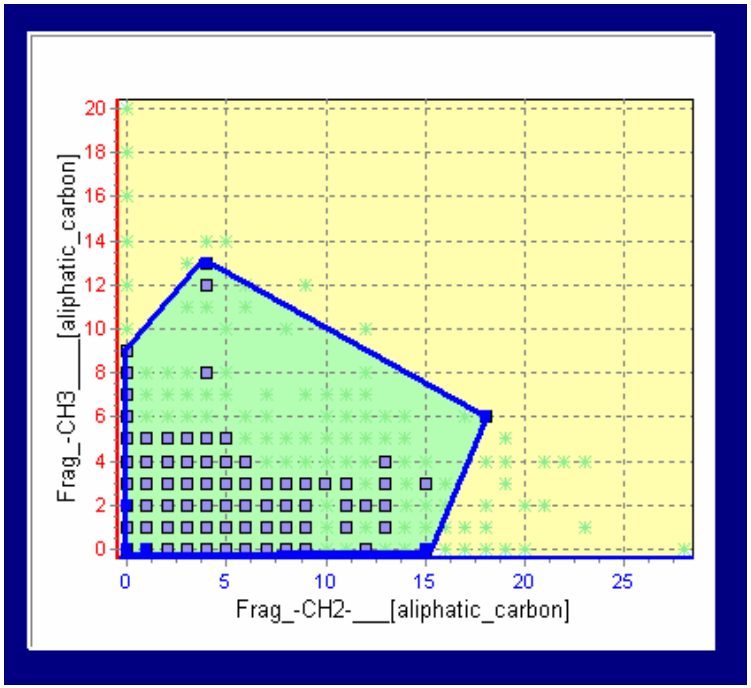
Хипер-правоъгълната област се дефинира в термините на главните компоненти:

$$[u^1_{\min}, u^1_{\max}] \times [u^2_{\min}, u^2_{\max}] \times \ldots \times [u^n_{\min}, u^n_{\max}]$$

- новата многомерна област е по-плътна с по-малки “дупки”.

- корелациите между променливите са избегнати

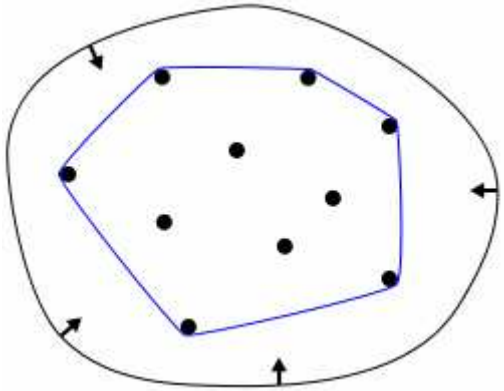
Геометричен метод



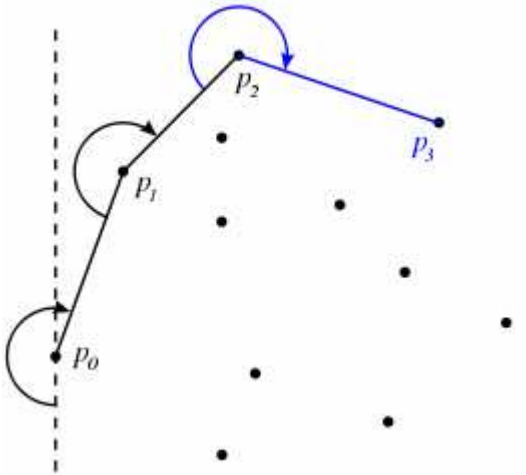
ОП (интерполационния регион) се пределя като **най-малката изпъкнала обвивка (МИО)** в многомерното пространство, която съдържа всички обекти от обучителната извадка.

Геометричен метод

Изчисляването на минималната изпъкнала обвивка (МИО) съдържаща дадено множество точки е изчислителна задача на аналитичната геометрия



- Съществуват ефективни алгоритми за определяна на МИО за 2D и 3D случаите
- При модели с висока размерност изчисляването на МИО е бавно и не ефективно



Характеристики на Минималната Изпъкнала Обвивка:

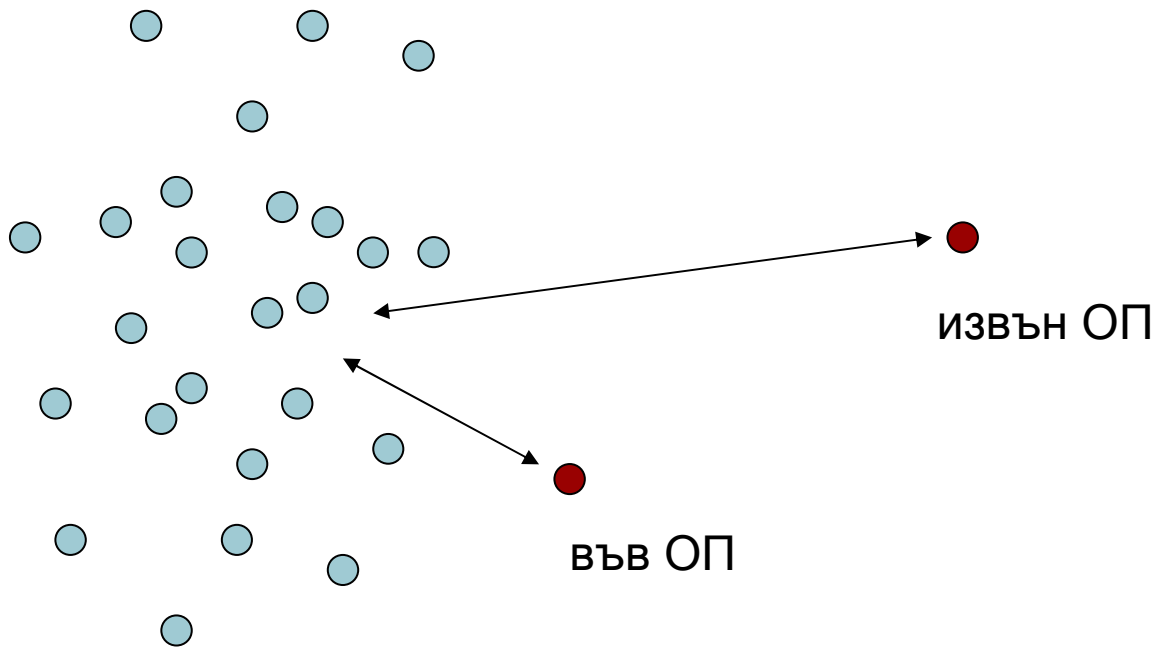
- Избегнати са големите празни пространства от периферията на хипер-правоъгълника (Вж. метода на интервалите)

- Съществуват области в МИО с по-гъсто разположени обекти и области с разредени обекти

- Не може да се определи местоположението на останалите празни области в МИО.

Метод на разстоянията

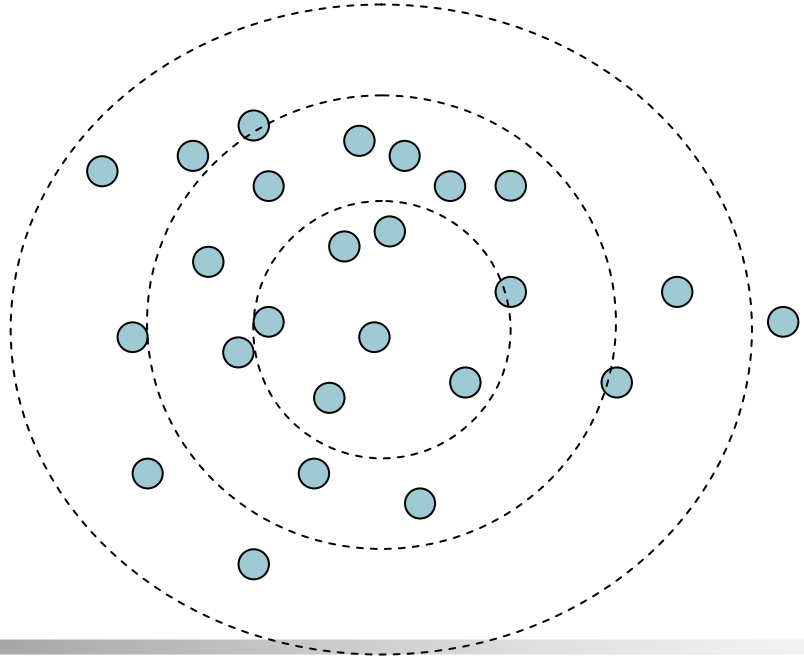
Изчислява се разстоянието от целевия обект до множеството обекти, които участват в обучителната извадка.



Метод на разстоянията

Решението дали даден обект е в ОП зависи от съществуването на критерий (прагова стойност) за разстоянието.

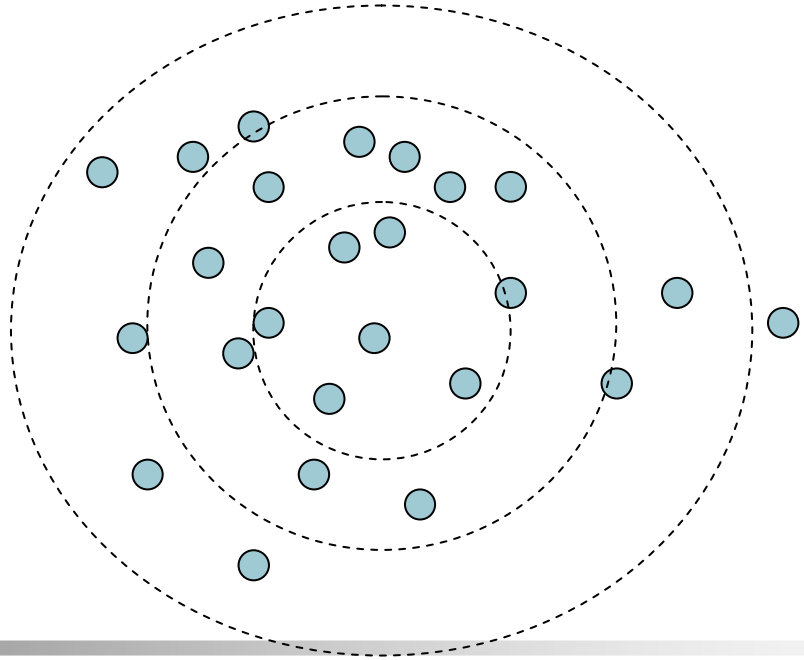
Регионите от химичното пространството, в които обектите са дадено константно разстояние оформят така-наречените **изо-контури**.



Метод на разстоянията

Чрез изо-контурите се оформят регион с предпола-гаема плътност на обектите.

Изо-контурните региони не отразяват реалната плътност на разпределение на обектите!

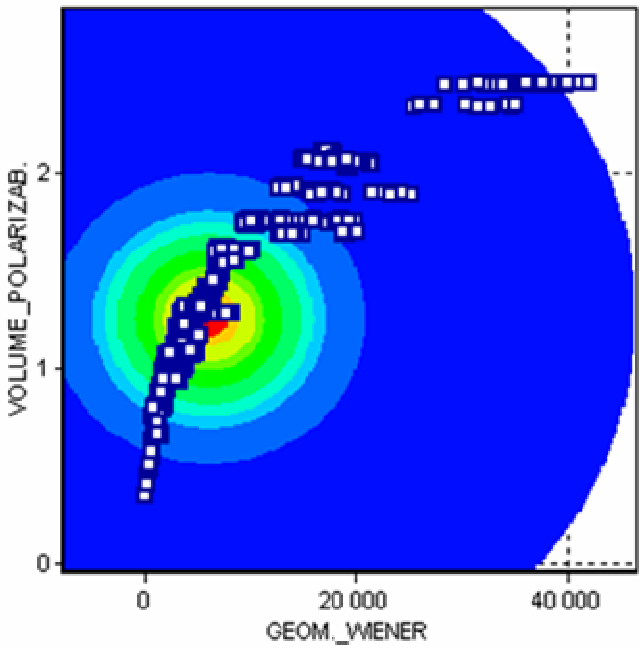


Метод на разстоянията

При използване на Евклидово разстояние се предполага:

- Гаусово разпределение на данните

- Липса на корелация между дескрипторите



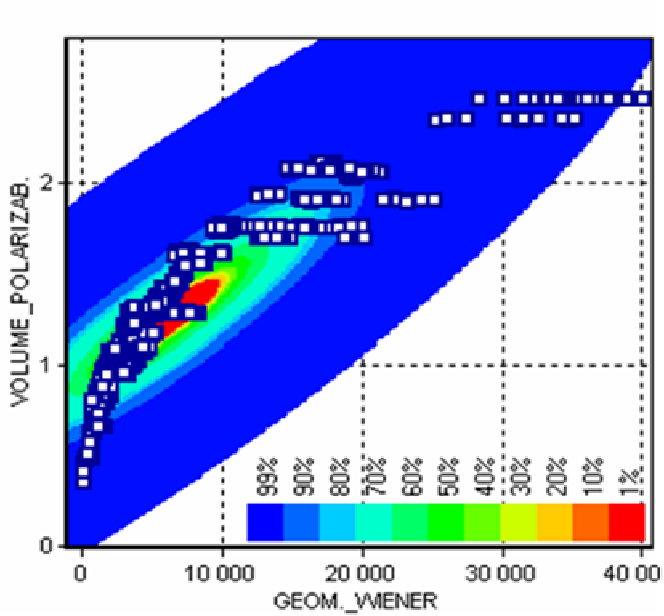
$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2}$$

Метод на разстоянията

При използване на разстояние на Махаланобис се предполага:

- Гаусово разпределение на данните

- Може да има корелация между дескрипторите



$$D_M(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}$$

Метод на разстоянията

Тест на Хотелинг (обобщена t-статистика) за многомерно пространство

За матрицата с дескрипторите ***X*** се изчислява така-наречената матрица на влиянието:

$$H = X(X^T X)^{-1} X^T$$

Диагоналните елементи на **H** (leverages) представляват мерки за разстоянието на даден обект от **X** до центроида.

Метод на разстоянията

Обектите по-близко до центроида на **X** имат по-малко влияние върху модела.

Чрез стойностите на **H** може да се открият обекти от обучителната извадка с мног силно влияние, които са много далече от центроида и представляват бегълци.

Метод на разстоянията

Уравнението за матрицата H може да се използва дефиниране на разстояние на произволен обект **z = (z_1, z_2, …,z_m)** до множеството обекти **X**

$$h(z) = z(X^T X)^{-1} z^T$$

Метод на разстоянията

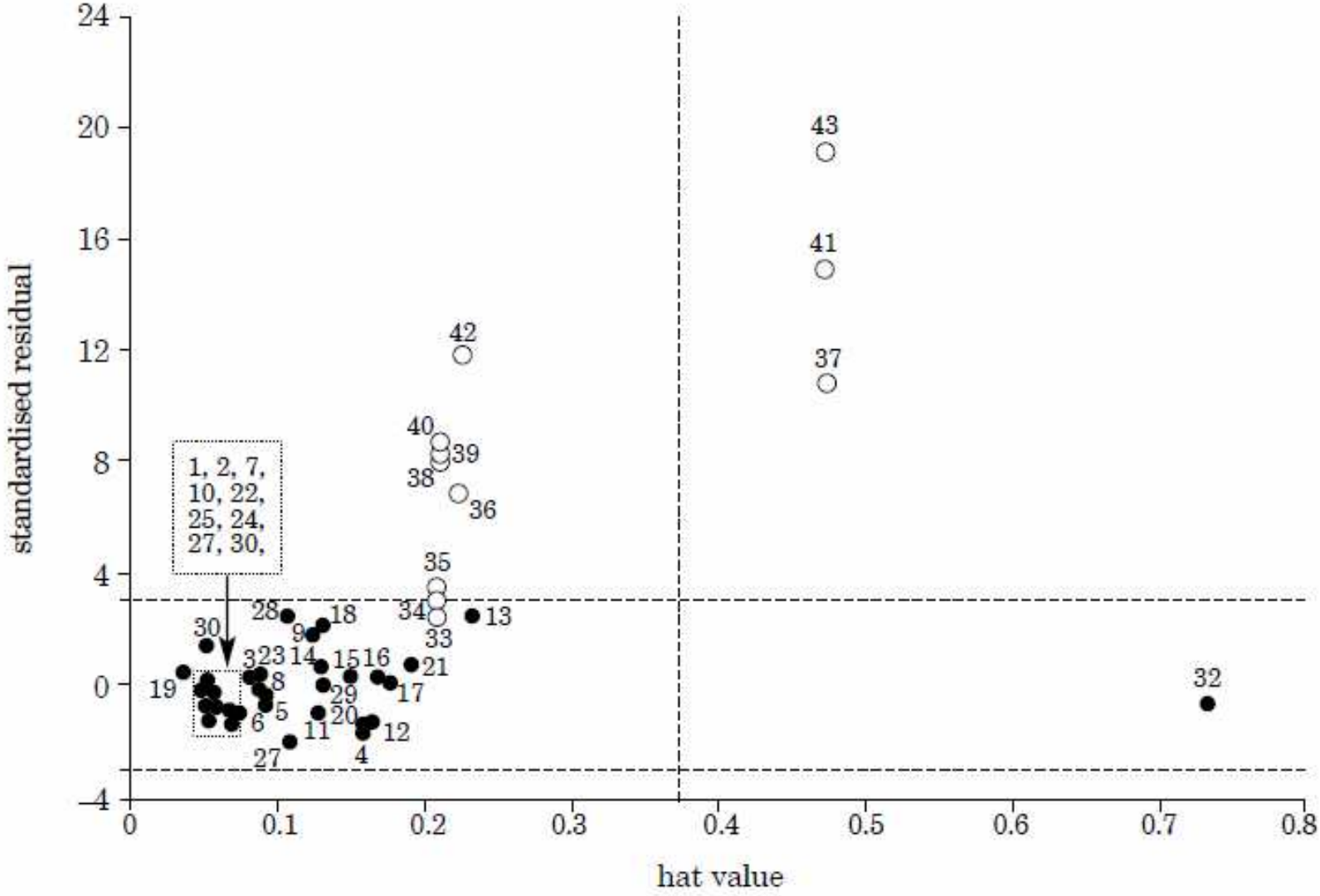
Статистика на Хотелинг за дефиниране на ОП. Обектът **z** е в областта на приложимост ако:

$$h(z) < h^*$$

Където критичната стойност **h*** (warning hat-value) се дефинира

$$h^* = \frac{3(m+1)}{n}$$

Метод на разстоянията



Литература:

1. Guidance Document On The Validation Of (Quantitative) structure-activity Relationships [(Q)sar] Models. OECD Environment Health and Safety Publications Series on Testing and Assessment No.69, ENV/JM/MONO(2007)2)

2. Tatiana I. Netzeva,et al. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure–Activity Relationships, ATLA 33, 1–19, 2005