QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review

Joanna Jaworska,[1] Nina Nikolova-Jeliazkova[2] and Tom Aldenberg[3]

[1]Procter & Gamble, Eurocor, Strombeek-Bever, Belgium; [2]Institute of Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria; [3]RIVM, Bilthoven, The Netherlands

Summary — As the use of Quantitative Structure Activity Relationship (QSAR) models for chemical management increases, the reliability of the predictions from such models is a matter of growing concern. The OECD QSAR Validation Principles recommend that a model should be used within its applicability domain (AD). The Setubal Workshop report provided conceptual guidance on defining a (Q)SAR AD, but it is difficult to use directly. The practical application of the AD concept requires an operational definition that permits the design of an automatic (computerised), quantitative procedure to determine a model's AD. An attempt is made to address this need, and methods and criteria for estimating AD through training set interpolation in descriptor space are reviewed. It is proposed that response space should be included in the training set representation. Thus, training set chemicals are points in n-dimensional descriptor space and m-dimensional model response space. Four major approaches for estimating interpolation regions in a multivariate space are reviewed and compared: range, distance, geometrical, and probability density distribution.

Key words: *applicability domain, multivariate interpolation, QSAR.*

Address for correspondence: *Joanna Jaworska, Procter & Gamble, Eurocor, Central Product Safety, 100 Temselaan, 1853 Strombeek-Bever, Belgium*.
E-mail: *Jaworska.j@pg.com*

Introduction

As the use of Quantitative Structure Activity Relationship (QSAR) models for chemical management increases, the reliability of the predictions from such models is a matter of growing concern. The OECD QSAR Validation Principles recommend that a model should be used within its applicability domain (AD; 1). The Setubal Workshop report (2) provides conceptual guidance on defining a (Q)SAR AD, but it is difficult to use directly. It states that: *The AD of a (Q)SAR is the physico-chemical, structural, or biological space, knowledge or information on which the training set of the model has been developed, and for which it is applicable to make predictions for new compounds. The AD of a (Q)SAR should be described in terms of the most relevant parameters i.e. usually those that are descriptors of the model. Ideally, the (Q)SAR should only be used to make predictions within that domain by interpolation not extrapolation.* This definition helps explain the intuitive meaning of the "AD" concept, but its practical application requires an operational definition allowing automatic (computerised) design and quantitative procedure to determine a model's AD. The lack of such guidance and tools for assessing ADs are discussed in a paper by Tunkel *et al*. (3).

Models yield reliable predictions when the models' assumptions are met, and unreliable predictions when these assumptions are violated. The chemical space occupied by a training data set is the basis for estimating where reliable predictions will occur, because, in general, interpolation is more reliable than extrapolation. A training set can be analysed in the model descriptor space, where chemicals are represented as points in a multivariate space, or directly by structural similarity analysis. The similarity approach to AD estimation relies on the premise that QSAR predictions are reliable if compounds are "similar" to the training set compounds (4). However, chemical similarity is a subjective term, and different concepts of similarity are relevant to different endpoints (5–8). This paper takes a statistical approach and examines AD assessment by estimation of interpolation regions in model descriptor space on the basis of the training data set.

Methods

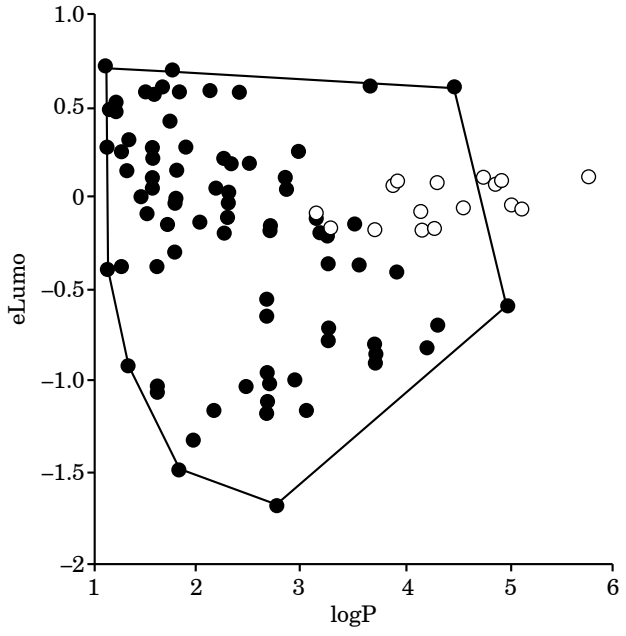
There are four major approaches to defining interpolation regions in multivariate space: range, distance, geometrical, and probability density distribution. Our choice of methods is not exhaustive, but focuses on the approaches which are most suitable as regression and classification models. Interpolation is the process of estimating values at arbitrary points between the points with known

values. An interpolation region in one-dimensional descriptor space is simply the interval between the minimum and maximum values of the training data set. A convex hull defines an interpolation region in multivariate descriptor space. The convex hull of a bounded subset of space is the smallest convex area that contains the original set. Readers seeking further details on the mathematical terms used in this paper may consult http://en.wikipedia.org/wiki/Main_Page.

Training sets in QSAR research

In practice, QSAR developers use retrospective data, often from different sources. Therefore, the selection of training data sets does not follow experimental design patterns; this results in large empty regions within the convex hull enclosing the data set. This paper's case study is an example of such a situation (9, 10; Figure 1). There are also chemicals outside the convex hull, but inside the ranges of the training set. The meaning of this is, in general, that different methods estimate the convex hull and therefore the AD in different ways. In order to provide guidance on choosing a method, we pay particular attention to the assumptions for each of the methods reviewed.

Figure 1: An eLUMO- logP 2D projection of training data set chemicals (9) and test set chemicals from Glende *et al*. (10)



● = *training data set chemicals;* ○ = *test set chemicals.*

Some authors consider the training set to consist only of independent variables (11, 12). Considering only the x-space part of the training set leaves out the information about the prediction space (i.e. y-space). Including both x-space and y-space allows the model to incorporate all the information contained in the training set, and prevents situations where data points are interpolated in x-space but are extrapolated in y-space. This can occur in linear models with two or more descriptors, with non-linear models with one or more descriptors, and is related to the amount of empty space included by a particular interpolation approach while estimating the AD. As the empty space a particular interpolation approach covers, declines, so does the need to include y-space in the assessment.

Thus, we propose to analyse the training set data in n-dimensional descriptor space *and* m (most often m = 1) dependent variables (property) space. In the case study analysed in this paper, we assess x-space and y-space separately. Because the usual dimensionality of y is 1, we assess the domains of y-values with a range method. A chemical will be in the domain if both conditions are satisfied. It is also possible to combine x-space and y-space and estimate joint interpolation space, but this approach is more suitable to probability distribution based methods.

Ranges

Descriptor ranges

The simplest method for approximating a convex hull is taking ranges of the individual descriptors. These ranges define an n-dimensional hyper-rectangle with sides parallel to the coordinate axes. The data are assumed to be distributed uniformly (13). The hyper-rectangle neither detects interior empty space nor corrects for correlations (linear or nonlinear) between descriptors. This approach may enclose considerable empty space, if the data are not uniformly distributed.

Principal components ranges

Principal Components Analysis (PCA) is a rotation of the data set to correct for correlations between descriptors. The principal components form a new orthogonal coordinate system. The rotation yields axes aligned with the directions of the greatest variations in the data set. The points between the minimum value and the maximum value of each principal component define an n-dimensional hyper-rectangle with sides parallel to the principal components. This hyper-rectangle AD includes empty space, but is smaller than the hyper-rectangle in the original descriptor ranges. This method has recently been used to analyse the KOWWIN model AD (14).

TOPKAT Optimal Prediction Space

The Optimum Prediction Space (OPS) from TOPKAT (11) uses a variation of PCA. As in typical PCA, data are centred, but around the average of each parameter range ($[x_{max}–x_{min}]/2$) instead of the standardised mean. The new orthogonal coordinate system is thus obtained (named the OPS coordinate system) by the same procedure — extracting eigenvalues and eigenvectors from the transformed data's covariance matrix. The minimum and maximum values of the data points on each axis of the OPS coordinate system define the OPS boundary. In addition, the Property Sensitive object Similarity (PSS) between the training set and a queried point assesses the confidence of the prediction. PSS is the TOPKAT implementation of a heuristic solution to reflect the data set's dense and sparse regions, and includes the response variable (y). Readers may consult the TOPKAT patent (11) for more details. A "similarity search" enables users to check the performance of TOPKAT in predicting the effects of a chemical that is structurally similar to the test structure. The user also has access to references to the original information sources.

Geometric methods

The direct method for estimating the coverage of an n-dimensional set is the convex hull calculation. Computing the convex hull is a computational geometry problem (15). Efficient algorithms for convex hull calculation are available for two and three dimensions; however, the algorithms' complexity rapidly increases in higher dimensions (for n points and d dimensions, the complexity is of order $O[n^{[d/2]+1}]$). This approach does not consider data distribution, but only analyses the set boundary. A convex hull cannot identify potential interior empty spaces.

Distance-based methods

Euclidean, Mahalanobis and City block distances

We review the three most useful distance methods in QSAR research: Euclidean, Mahalanobis, and City block distances. Distance-based approaches calculate the distance from each point to a particular point in the data set. Distance to the mean, averaged distance between the query point and all points in the data set, and maximum distance between the query point and data set points, are examples of the many available options. Categorising data points as close to/in the data set depends on the threshold chosen by the user.

Euclidean and Mahalanobis distance methods identify the interpolation regions by assuming that the data are normally distributed (13, 16). City-block distance assumes a triangular distribution. Mahalanobis distance is unique, because it automatically takes into account the correlation between descriptor axes through a covariance matrix. Other approaches require the additional step of PC rotation to correct for correlated axes. City block distance is particularly useful for discrete descriptors. The shape of the iso-distance contours (for example, the regions at a constant distance) depends on the particular distance measure used (see Table 1) and on the particular approach for measuring the distance between a point and a data set.

Hotelling T^2 and leverage

Hotelling T^2 test and leverage, also distance methods, have been recommended for assessing QSAR ADs (17, 18). These measures are proportional to each other and to the Mahalanobis distance. The Hotelling T^2 method is a multivariate Student's *t*-test and assumes a normal data distribution, as does the leverage approach (19). In regression, the term "leverage values" refers to the diagonal elements of the hat matrix $H = (X(X'X)^{-1}X')$. A given diagonal element ($h[_{ii}]$) represents the distance between the X value for the ith observation and the means of all X values. These values indicate whether X values may be outliers (16, 20). Both Hotelling T^2 and leverage correct for colinear descriptors through use of the covariance matrix.

Hotelling T^2 and leverage measure the distance of an observation from the centre of a set of X observations. A tolerance volume is derived for Hotelling T^2 (18). For leverage, a value of 3 is commonly used as a cut-off value for accepting predictions, because points that lie +/– 3 standard deviations from the mean cover 99% of normally distributed data.

High leverage values do not always indicate outliers for the model, i.e. points that are outside the model domain. If high leverage points fit the model well (i.e. have small residuals), they are called "good high leverage points" or good influence points. Such points stabilise the model and make it more precise. High leverage points, which do not fit the model (i.e. have large residuals) are called "bad high leverage points" or bad influence points. The field of robust regression provides a number of methods for overcoming the sensitivity of Hotelling T^2 and leverage to unusual observations, but that is beyond the scope of this paper.

Probability density distribution methods

Another approach to estimating the interpolation region is the use of probability density distribution

Table 1: List of data distribution assumptions for distance approaches

Distance measure	Assumption on data distribution	Shape of contour lines
Mahalanobis/Hotelling T^2/leverage $D_M(x,\mu) = (x-\mu)^T \Sigma^{-1}(x-\mu)$ covariance matrix Σ	Multivariate normal, mean μ, covariance matrix Σ $p(x) = \frac{1}{(2\pi)^{N/2} \lvert\Sigma\rvert^{1/2}} exp\{-\frac{1}{2} D_M(x,\mu)\}$	Ellipses (hyper-ellipses), defined by matrix Σ
Euclidean $D_E(x,\mu) = \sqrt{(x-\mu)^T (x-\mu)}$	Multivariate normal, mean μ, unit covariance matrix $p(x) = \frac{1}{(2\pi)^{N/2}} exp\{-\frac{1}{2} D_E^2(x,\mu)\}$	Spheres (hyper-spheres)
City block $d(x,y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$	Multivariate uniform	Rectangle with edges that must be traversed to get from point a to b within the grid, identical to getting from corner a to b in a rectilinear downtown area, hence the name "city-block metric"

methods (21). Parametric and non-parametric methods are the two major approaches. Parametric methods use the probability density function p(x) of standard distributions such as Gaussian and Poisson distributions. Alternatively, non-parametric techniques permit the estimation of probability density solely from data.

Non-parametric probability density estimation is free of assumptions about the data distribution, and is often referred to as a distribution-free method. It is the only approach capable of identifying internal empty regions within the convex hull. Furthermore, if empty regions are close to the convex hull border, non-parametric methods can generate concave regions to reflect the actual distribution of the data. In other words, this method captures the actual data distribution. Finally, there is no need to specify a reference point in the data set. Instead, a probability of belonging to the set is calculated for each data point. Because of these attractive features, which are lacking in other estimation methods, probability density estimation is explained in more detail in Appendix 1.

Relationships between different interpolation approaches

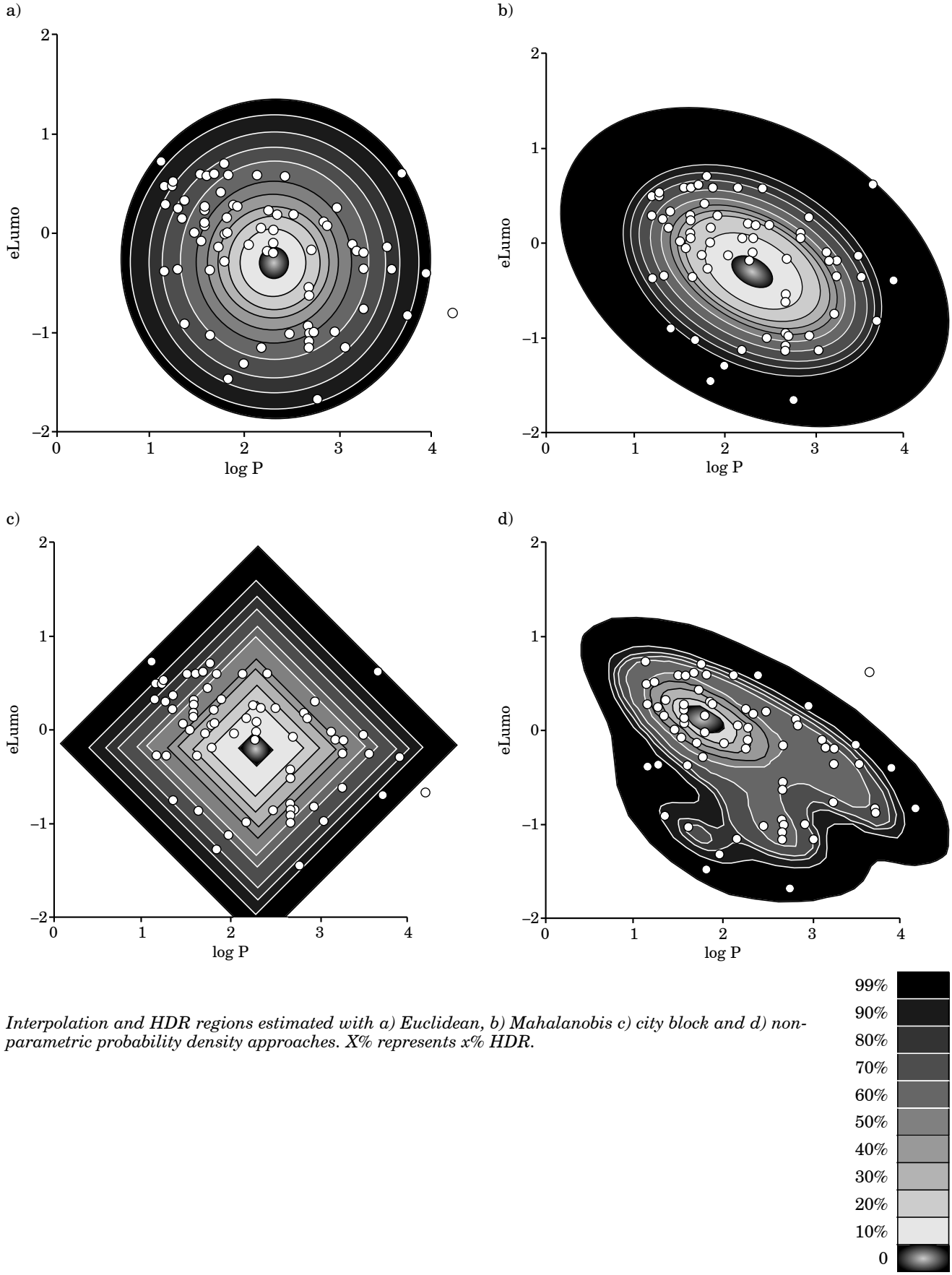
Probability density and all distance-based methods yield proportional results if the data are normally distributed (Table 1), and the particular distance method uses the data mean as a point of reference. For all other distributions, the distance values are not proportional to probability density, nor do they identify the presence of dense and empty regions. There is no general rule on which methods will yield the most different results; it all depends on the specific data distribution. For comparison, Figure 2 displays AD regions estimated by Euclidean, Mahalanobis and City block distances, and probability density distribution based on the case study data set (9).

Table 2 summarises the reviewed methods for assessing interpolation regions. In addition to technical aspects, we compare software availability for each method. The complexity of these approaches varies greatly. This may turn some users away from the more complex but more flexible approaches. However, increasing accessibility to sophisticated numerical methods through software packages allows even non-experts to apply computationally difficult methods.

Different results versus different methods

Figure 2 shows that different interpolation approaches yield different ADs. This may leave readers wondering which method to choose in specific situations. The choice of the particular method is straightforward: the data distribution must meet the assumptions of the method. If the training set data are uniformly distributed, the ranges approach would be recommended. If not, data distribution should be tested for normality; if normality is confirmed, AD can be estimated by one of the distance approaches. If the normality test fails, AD assessment based on other parametric distributions or a non-parametric probability distribution should be considered.

Figure 2: Interpolation regions and respective highest density regions (HDR) of the training set (9)



Interpolation and HDR regions estimated with a) Euclidean, b) Mahalanobis c) city block and d) non-parametric probability density approaches. X% represents x% HDR.

Table 2: Characteristics of the reviewed interpolation methods

Criterion	Ranges	OPS	Euclidean distance	Mahalanobis distance/ Hotelling T^2/leverage	Geometric	Probability density via non-parametric kernel estimation
Assumption regarding data distribution	Uniform	Uniform	Normal, equal variances	Normal, arbitrary variances	Arbitrary distributions	None, reflects actual distribution of any data set
Assumption regarding model descriptors	Uncorrelated, PC rotation can be added as a pretreatment step	Arbitrary (uses PC rotation)	Uncorrelated descriptors	Arbitrary descriptors, potential correlations are accounted for directly in the formula	Uncorrelated, additional step of PC rotation can be added	PC rotation is necessary as a pre-treatment of the data step
Ability to discover internal dense and sparse regions of the interpolated space	No	No	No	No	No	Yes
Ability to quantify distance from the centre of the set	No	Yes	Yes	Yes	No	Yes
Ease of application of the method in many dimensions	Easy	Easy	Easy	Easy, however involves inversion of the covariance matrix (could be slow for many dimensions)	Difficult above 3D	Used to be difficult above 3D, a recent very fast method in Matlab works in many dimensions
Availability of tools i.e. software using the method	Statistical software for general use	TOPKAT	Statistical software for general use	Statistical software for general use, may require some programming	Computational geometry packages; most mathematical packages	Matlab, Mathematica HDR calculation requires programming

OPS = optimum prediction space, PC = principal components, HDR = highest density region.

Although it is straightforward, the requirements for the data distribution in a training set are becoming quite complex. First, the data must conform to the assumptions of a particular model fitting technique. Second, the data must also meet the assumptions of the domain estimation method. As an example, we use the linear regression model as a modelling technique and estimate the AD by leverage.

When fitting data to a linear regression model with Ordinary Least Squares, one first needs to determine whether the following conditions hold: 1) the residuals are normally distributed; 2) the error distribution has a mean of zero; 3) the variance of the random error is constant for all values of x; and 4) the errors associated with any two observations are independent. That is, the error associated with one value of y has no effect on the errors associated with other values. Assumptions 1 and 2 can be checked by residual analysis. Verifying assumption 3 requires information about the experimental data measurement error. Different experimental tests may have different variances; researchers must be alert for this when combining data from different tests. While assumption 4 is usually satisfied in QSAR research, very few papers report whether assumptions 1–3 are satisfied, suggesting that this step may be frequently overlooked.

Next, modellers need to choose a method for AD estimation that is suitable for the data distribution. Let us consider the leverage method. This is an appealing approach, because the hat matrix needed to identify high leverage points (here identified as out of the domain) is automatically calculated dur-

ing regression model development. However, this method is only suitable when the data are normally distributed. Note that in the model development phase, developers check *error* distribution not *data* distribution.

In the past, verifying whether a particular training set was appropriate for a particular modelling technique was rare in QSAR research. Instead, tradition and convenience determined the choices. Now, in order to develop ADs effectively, we need to re-examine existing training sets. It should not be surprising that complexity of the training set distribution will be matched by the complexity of the *appropriate* AD estimation method.

The AD estimation process may expose some deficiencies in existing models. For example, if the model uses collinear descriptors, it should be redeveloped, so the descriptors are orthogonal instead of correcting for the collinearity during AD estimation (14). Running PCA rotation only during AD estimation may make it difficult to interpret predictions, as the AD will be projected in a space different from the original model descriptor space. One possible solution is to start developing modelling approaches that will allow for simultaneous model development and AD estimation. Simultaneous development will avoid imposing double, often different, requirements for the data distribution related to model development and to AD estimation.

Case Study: AD of the Salmonella Mutagenicity of Aromatic Amines QSAR Model

The model

Debnath *et al*. (9) published a mutagenicity model for $n = 88$ aromatic and heteroaromatic amines: $\log TA98 = 1.08 \log P + 1.28eHOMO - 0.73eLUMO + 1.46 IL + 7.20$ (1), where log P is the n-octanol/water partition coefficient, eHOMO and eLUMO are energy on the highest occupied and lowest unoccupied molecular orbitals, respectively, and IL is an indicator variable with a value of 1 for compounds containing three or more fused rings and 0 for all other species.

Glende *et al.* (10) studied 18 alkyl-substituted (ortho to the amino function) derivatives not included in the original data of Debnath *et al.* (9). Most of these new chemicals had descriptor values in the range of the original chemicals. However, with growing steric hindrance of the alkyl groups, the difference between the predicted and experimental values increased. Glende *et al.* concluded that the QSAR equation is not appropriate for evaluating the mutagenicity of aromatic amines substituted with such alkyl groups. The set of Glende *et al.* was used as the test set in the case study.

The methods for AD estimation

We assessed the AD of the model of Debnath *et al.* (9) by using the following approaches:

1. Ranges in descriptor space and in PC rotated space;
2. Euclidean distance in descriptor space and in PC rotated space;
3. City-block distance in descriptor space and in PC rotated space;
4. Hotelling T^2;
5. Probability density distribution; and
6. Range of the response variable.

We developed criteria appropriate to each method for in the domain and out of the domain. For the ranges method, a chemical is out of domain if at least one descriptor is out of range or a combination of descriptors are out of range (this is equivalent to the endpoint value being out of range). For Hotelling T^2 and distance methods, the cut-off threshold was the largest distance among the training set points to the centre of the training data set. For probabilistic density, the cut-off thresholds were the 95th and 99th percentiles of the training set's probability density.

The training set data distribution failed the Kolmogorov–Smirov uniform distribution tests and Jarque–Bera normality tests implemented in MATLAB 6.5 R13 at $p = 0.05$. This suggests that only non-parametric probability distribution estimation methods are suitable for the model (9). Nevertheless, for comparison purposes, we carried out AD estimation with all five methods.

Figures 3–5 illustrate the correspondence between domain assessment and prediction error for the approaches evaluated. There is a trend indicating that the average prediction error for chemicals in the domain is smaller than that for chemicals out of the domain for all the methods. In that sense, the results are similar to the findings of Tong *et al*. (12). This observation is trivial, but it is interesting to see the quantitative difference in the quality prediction in and out of the domain (Table 3).

Table 3 summarises the results of the AD estimation approaches evaluated. The numbers of chemicals in and out of the domain, their identification numbers, and the root mean square errors (RMSEs) for chemicals in and out of the domain are included. The RMSE is a sum of squared prediction errors, divided by the number of points. By using the RMSE, we do not question the goodness of fit of analysed QSAR model analysed, but rather use it as

Table 3: Summary statistics for different approaches to application domain assessment

Domain defined by:	PC rotation + scaling	Validation (in)			Validation (out)		
		No.	RMSE	No. compounds	No.	RMSE	No. compounds
Ranges		15	1.9557	All but 9, 10, 18	3	3.608	9, 10, 18
Euclidean distance		15	1.9557	All but 9, 10, 18	3	3.608	9, 10, 18
City block distance		17	2.0708	All but 18	1	4.85	18
Hotelling T^2		17	2.0708	All but 18	1	4.85	18
Probability density 95%		8	1.26	1, 2, 3, 6, 7, 11, 12, 16	10	2.9	4, 5, 8, 9, 10, 13, 14, 15, 17, 18,
Probability density 99%		12	1.7	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 16	9	3.2	9, 10, 14, 15, 17, 18,
Ranges	Yes	17	2.0708	All but 18	1	4.85	18
Euclidean distance	Yes	15	1.9557	All but 9, 10, 18	3	3.608	9, 10, 18
City block distance	Yes	13	1.9403	All but 9, 14, 15, 17, 18	5	3.0816	9, 14, 15, 17, 18
Probability density 95%	Yes	8	1.26	1, 2, 3, 6, 7, 11, 12, 16	15	2.5061	4, 5, 8, 9, 10, 13, 14, 15, 17, 18
Probability density 99%	yes	9	1.7	1, 2, 3, 5, 6, 7, 11, 12, 16	9	2.8	4, 8, 9, 10, 13, 14, 15, 17, 18

The numbering of compounds is as in Glende (8).
PC = principal components, RMSE = root mean square error.

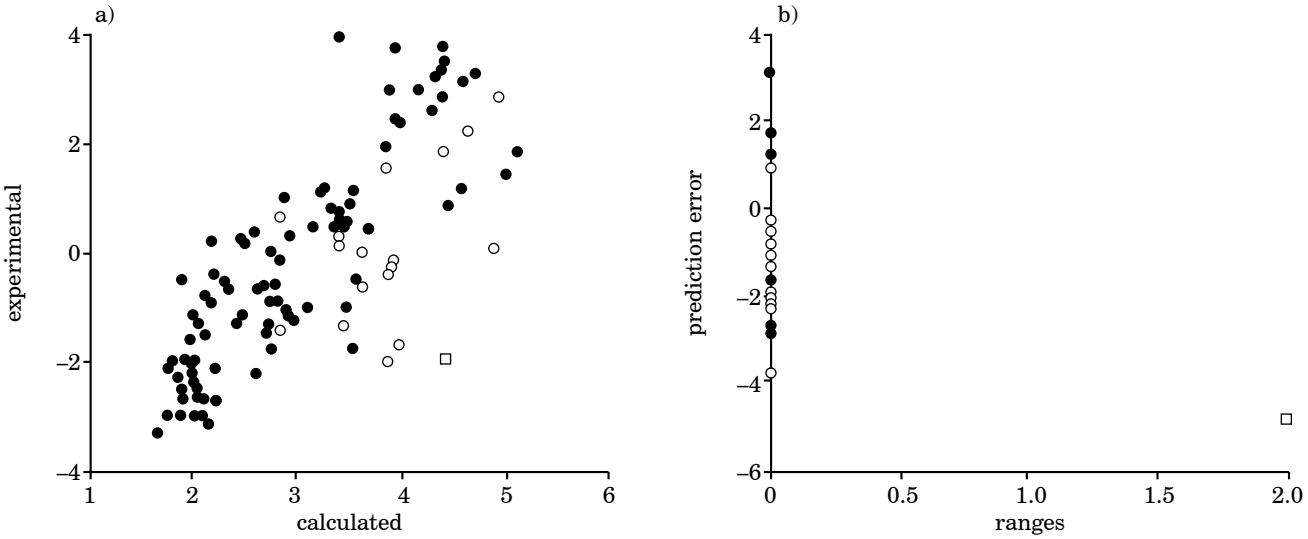
a relative measure of prediction accuracy in the domain and out of the domain. The RMSE of out of domain validation points exceeds the RMSE of all validation points for all approaches. As expected, the RMSE for the chemicals in the domain estimated by probability density approach is the lowest among the methods considered; this confirms that this method is the most accurate and appropriate. In this case study, ranges of the response variable were always in the domain and did not influence the results.

Discussion

We reviewed AD estimation methods by determining interpolation regions defined by the training data set in model descriptor space. By focusing on the training set, we did not discuss a particular QSAR modelling approach. However, we would like to stress that our discussion is more suited to the low-dimension regression and classification models prevalent in modelling safety endpoints (22). The discussion is less appropriate to the partial least squares approach, which has its own set of diagnostic tools, such as distances to the model in x and y space, DMODX and DMODY, respectively (18). However, similarly to the situation with regard to the partial least squares approach, DMODY, we emphasise the need to include y-space in the AD estimation, particularly for approximating training set coverage.

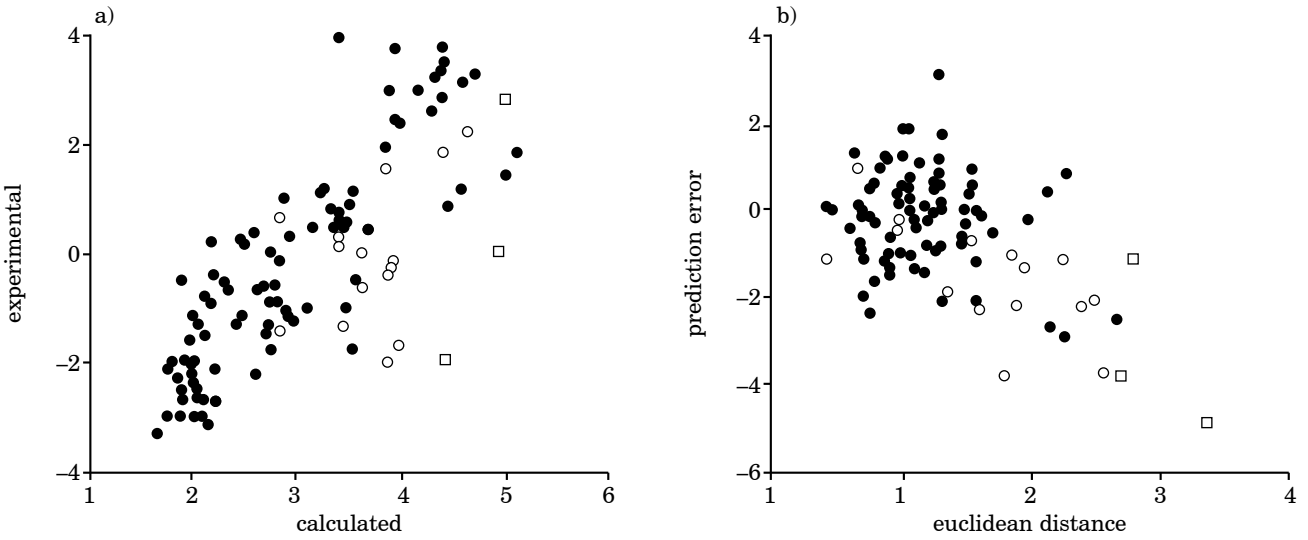
The results of AD estimation on the basis of training set coverage in descriptor space reveal a general trend — interpolative predictive accuracy, i.e. concordance between observed and predicted values, was, on average, greater than extrapolative predictive accuracy. That, however, is only true on average — many compounds with small errors are outside the training set coverage, as there are also compounds with large errors inside the domain.

Figure 3: Correspondence between the prediction error and the application domain boundary obtained by principal components rotated ranges



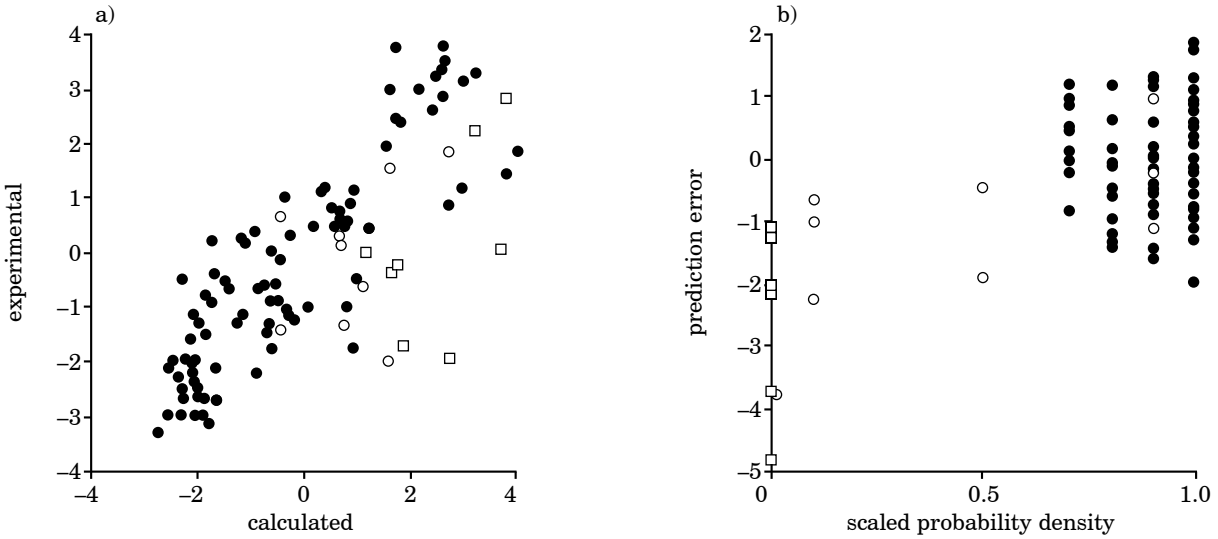
In the domain: training set (9) (●) and test set (10) (○); out of the domain: training set (■) and test set (□).
a) experimental vs. calculated values in original space; b) correspondence between prediction error and the number of dimensions where the point is out of training set range (i.e. zero means in-range).

Figure 4: Correspondence between the prediction error and the application domain boundary obtained by the Euclidean distance



In the domain: training set (8) (●) and test set (4) (○); out of the domain: training set (■) and test set (□).
a) experimental vs. calculated values in original space; b) correspondence between prediction error and the distance between a point and the mean for the test set.

Figure 5: Correspondence between the prediction error and the application domain boundary obtained by the non-parametric probability density method



In the domain: training set (9) (●) and test set (10) (○); out of the domain: training set (■) and test set (□). a) experimental vs. calculated values in original space; b) correspondence between prediction error and probability density value. The 99th percentile was used as the cut-off value, points with a probability value less then 0.01 are deemed as out of the domain.

Different interpolation approaches yield different ADs. By emphasising the need to meet each interpolation method's assumptions, we demonstrate that analysing the training data set distribution provides clear guidance on choosing a particular method (as the methods are mutually exclusive). However, in practice, it is not always possible to choose the correct method by considering only data distribution. The dimensionality of the model needs to be considered as well, because the number of data points in the training set may be insufficient for the application of a particular approach (14). High model dimensionality increases numerical complexity, especially for geometric and non-parametric probability density approaches (21).

In addition, researchers need to reconcile the methods used for model development (fit) and the methods used for estimating the AD. Treating these two steps separately often imposes different requirements for the data distribution in the training set, making it difficult to meet all the assumptions. Joint fit and estimation of AD by probability density methods is a promising approach. This area requires more attention and further work.

By identifying the training data set coverage in descriptor space, we make only a partial step toward defining a model's AD. There is always a possibility that the model lacks a descriptor needed to correctly predict the activity of a chemical. Thus, despite the fact that the chemical appears to be in the descriptor domain, its activity will most likely be predicted with error, because it is structurally different from the training set. There is also another possibility — that the model extrapolates correctly outside the domain.

Therefore, to describe the domain more robustly, the full training set comprising both structures and descriptor set is required. The full training set permits an assessment of its chemical space coverage. In this paper, we have not discussed the need for a global structural similarity test to ensure that the structural features in a new test compound are covered in the original training set of chemicals (a quantitative measure of uniqueness relative to the training set; Dave Stanton, personal communication). The global similarity test should be sufficiently robust to cover general chemistry. Then, the AD estimation would consist of two steps: 1) training set coverage in terms of a descriptor values assessment; and 2) structural similarity identification. We recommend using different methodologies to examine the training set in different ways, in order to maximise the chances of finding a potential difference. These ideas were recently discussed at an ECVAM Workshop on the AD (23), and our review was used as the background paper for the workshop. The methods described in this paper are implemented in the software AMBIT, developed by us and available free from http://ambit.acad.bg/.

Lastly, we would like to stress that if a chemical is inside the domain according to a given, correctly applied method, this is not a final argument for accepting the prediction; rather, it is an indication of the correct application of a model and the *reduced* uncertainty of a prediction. This uncertainty can be expressed as the RMSE, confidence intervals (16) or other methods. Similarly, if a chemical is outside the domain according to a given, correctly applied method, this is not a final argument for rejecting the prediction; rather, it is an indication of the *increased* uncertainty of the prediction. We can say that this is, in a statistical sense, an incorrect application of a model, but it is nevertheless possible that the model will generate a correct result.

Acknowledgments

Funding for Nina Nikolova-Jeliazkowa was provided by Procter & Gamble as a postdoctoral fellowship. All the authors acknowledge EU JRC ECVAM for partial funding of this work under contract CCR.496575-Z. We thank Dr. A. Benigni for drawing our attention to the selected case study. The AMBIT software development was funded by the CEFIC Long Research Initiative (project EEM-9).

Received 27.1.05; received in final form 7.6.05; accepted for publication 13.6.05.

References

1. OECD (2004). *Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models*. Paris, France: OECD. Website http://www.oecd.org/document/23/0,2340, en_2649_201185_33957015_1_1_1_1,00.html (Accessed 27.05.05).
2. Jaworska, J., Comber, M., Van Leeuwen, C. & Auer, C. (2003). Summary of the workshop on regulatory acceptance of QSARs. *Environmental Health Perspectives* **111**, 1358–1360.
3. Tunkel, J., Mayo, K., Austin, C., Hickerson, A. & Howard, P. (2005). Practical considerations on the use of predictive models for regulatory purposes. *Environmental Science and Technology* **39**, 2188–2199.
4. Sheridan, R., Feuston, R.P., Maiorov, V.N. & Kearsley, S. (2004). Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Science* **44**, 1912–1928.
5. Nikolova, N. & Jaworska, J. (2002). Review of chemical similarity measures. *QSAR & Combinatorial Science* **22**, 1006–1026.
6. Bender, A. & Glen, R.C. (2004). Molecular similarity: a key technique in molecular informatics. *Organic and Biomolecular Chemistry* **2**, 3204–3218.
7. Kubinyi, H. (1998). Similarity and dissimilarity: A medicinal chemist's view. *Perspectives in Drug Discovery and Design* **9–11**, 225–252.
8. Martin, Y.C., Kofron, J.L. & Traphagen, L.M. (2002). Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry* **45**, 4350–4358.
9. Debnath, A.K., Debnath, G., Shusterman, A.J. & Hansch, C. (1992). A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Environmental Molecular Mutagenesis* **19**, 37–52.
10. Glende, C., Schmitt, H., Erdinger, L., Engelhardt, G. & Boche, G. (2001). Transformation of mutagenic aromatic amines into non-mutagenic species by alkyl substituents. Part I. Alkylation ortho to the amino function. *Mutation Research* **498**, 19–37.
11. TOPKAT OPS (2000). US patent no. 6 036 349 issued March 14, 2000.
12. Tong, W., Qian, X., Huixiao, H., Leming, S., Hong, F. & Perkins, R. (2004). Assessment of prediction confidence and domain extrapolation of two Structure–Activity Relationship models for predicting estrogen receptor binding activity. *Environmental Health Perspectives* **112**, 1249–1254.
13. Seber, G.A.F. (1984). *Multivariate Observations*. 712pp, New York, NY, USA: John Wiley & Sons.
14. Nikolova-Jeliazkova, N. & Jaworska, J. (2005). An approach to determining AD for QSAR group contribution models: an analysis of SRC KOWWIN. *ATLA* **33**, 461–470.
15. Preparata, F.P. & Shamos, M.I. (1991). Convex hulls: Basic algorithms. In *Computational Geometry: An Introduction* (ed. F.P. Preparata & M.I. Shamos), pp. 95–148, New York, NY, USA: Springer-Verlag.
16. Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. 2nd edn, 592pp. Computer Science and Scientific Computing Series. New York, NY, USA: Academic Press.
17. Tropsha, A., Gramatica, P. & Gombar, V. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR Models. *QSAR & Combinatorial Sciences* **22**, 69–77.
18. Eriksson, L., Jaworska, J., Worth, A., Cronin, M.T.D., McDowell, R.M. & Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environmental Health Perspectives* **111**, 1351–1375.
19. Neter, J., Kutner, M.H., Wasserman, W. & Nachtsheim, C. (1996). *Applied Linear Statistical Models*. 1408pp. New York, NY, USA: McGraw-Hill.
20. Myers, R.H. (2000). *Classical and Modern Regression with Applications*. 2nd edn, 488pp. Pacific Grove, CA, USA: Duxbury Press.
21. Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. 176pp. Boca Raton, FL, USA: CRC Press.
22. ECETOC. (2003). Evaluation of commercially available software for human health, environmental and physico-chemical endpoints. *ECETOC QSAR TF report nr 89*, 164pp. Brussels, Belgium: ECETOC.
23. Netzeva, T.N., Worth, A.P., Aldenberg, T., Benigni, A., Cronin, M.T.D., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D., Schultz, T.W., Stanton, D., van de Sandt, J.J., Tong, W., Veith, G. & Yang, C. (2005). Current status

of methods for defining the applicability domain of (Quantitative) Structure-Activity Relationships. The report and recommendations of ECVAM Workshop 52. *ATLA* **33**, 1–19.
24. Gray, A. & Moore, A. (2003). Nonparametric density estimation: toward computational tractability. Website http://www.cs.cmu.edu/~agray/nbody.html (Acccessed 12.09.05).
25. Gray, A.G., Moore, A.W., Nichol, R.C., Connolly, A.J., Genovese, C. & Wasserman, L. (2003). Very fast multivariate kernel density estimation via computational geometry. Website http://www.adass.org/adass/proceedings/adass03/O3-2/ (Accessed 14.09.05)
26. Ihler, A. (2004). *MATLAB KDE Class*. Website http://ssg.mit.edu/~ihler/code/kde.shtml (Accessed 27.01.05).
27. Friedman, J.H. (1987). Exploratory projection pursuit. *Journal of the American Statistic Association* **82**, 249–266.
28. Chen, M.-H. & Shao, Q-M. (1999). Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistics* **8**, 69–92.
29. Weinzierl, S. (2000). Introduction to Monte Carlo Methods. Website http://arxiv.org/abs/hep-ph/0006269 (Accessed 27.01.05).
30. Press, W., Teukolsky, S., Vetterling, W. & Flannery, B. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edn, 1020pp. Cambridge, UK: Cambridge University Press.

Appendix 1

Probability density estimation

Density estimation is an area of extensive research. Due to computational challenges, most methods focus on low dimensional (1-, 2-, 3-) densities, unless additional assumptions are made (24). The recently developed Algorithm for Multivariate Kernel Density Estimation (24–26) achieves several orders of magnitude speed improvement by using computational geometry to organise the data. This method directly estimates true multivariate density, and is very accurate.

The kernel density estimation method

In the kernel probability density estimation of m-dimensional descriptor space, m-dimensional kernels are placed on every data point and then are summed up. With many data points and a high-dimensional descriptor space, this procedure requires considerable time and computer memory, consuming calculation. Estimating the joint probability as a product of marginal (one-dimensional) probabilities compromises the quality of the estimate, if descriptors are statistically dependent (27).

$$p(x_1,\ldots,x_n) = \prod_{k=1}^{n} p(x_k) \quad (2)$$

Lack of dependence is a much stronger requirement than lack of linear correlation. It means lack of both linear and any nonlinear correlation between the descriptors. Descriptor independence is rare in real data sets. While it is difficult to account for every possible nonlinear correlation, linear correlation is easy to handle via PCA.

There are four important steps in probability density estimation:

1. Standardising the data (scale and centre);
2. Extracting the principal components of the data set;
3. Skewness correction transformation along each principal component; and
4. Estimating the one-dimensional density on each transformed principal component.

Figure 6 illustrates three projections of probability density with increasing accuracy. The density obtained as the product of 1-D densities in the original descriptor space is shown in Figure 6a. The estimated density does not reflect the actual data density, because the parameters are dependent. Figure 6b displays the density obtained as a sum of Gaussian kernels in PC space. Figure 6c shows the improved quality of the estimated density obtained by employing data set transformations such as standardising the data and correcting for skewness. Figure 6 illustrates that there is a need to be very transparent about data processing during density estimation, as different results can be obtained.

Mathematical details of highest density region calculation

The next step after probability density estimation is to find the highest density regions (HDR) which comprise a predefined fraction of the total probability mass. The $(1 - \alpha)$ HDR region is the smallest interval (in 1-D) or multidimensional region (> 1-D), comprising $(1 - \alpha)*100$ percents of the probability mass, where $(0 < \alpha < 1)$ (Figure 7). The user can choose different α levels for the AD boundary.

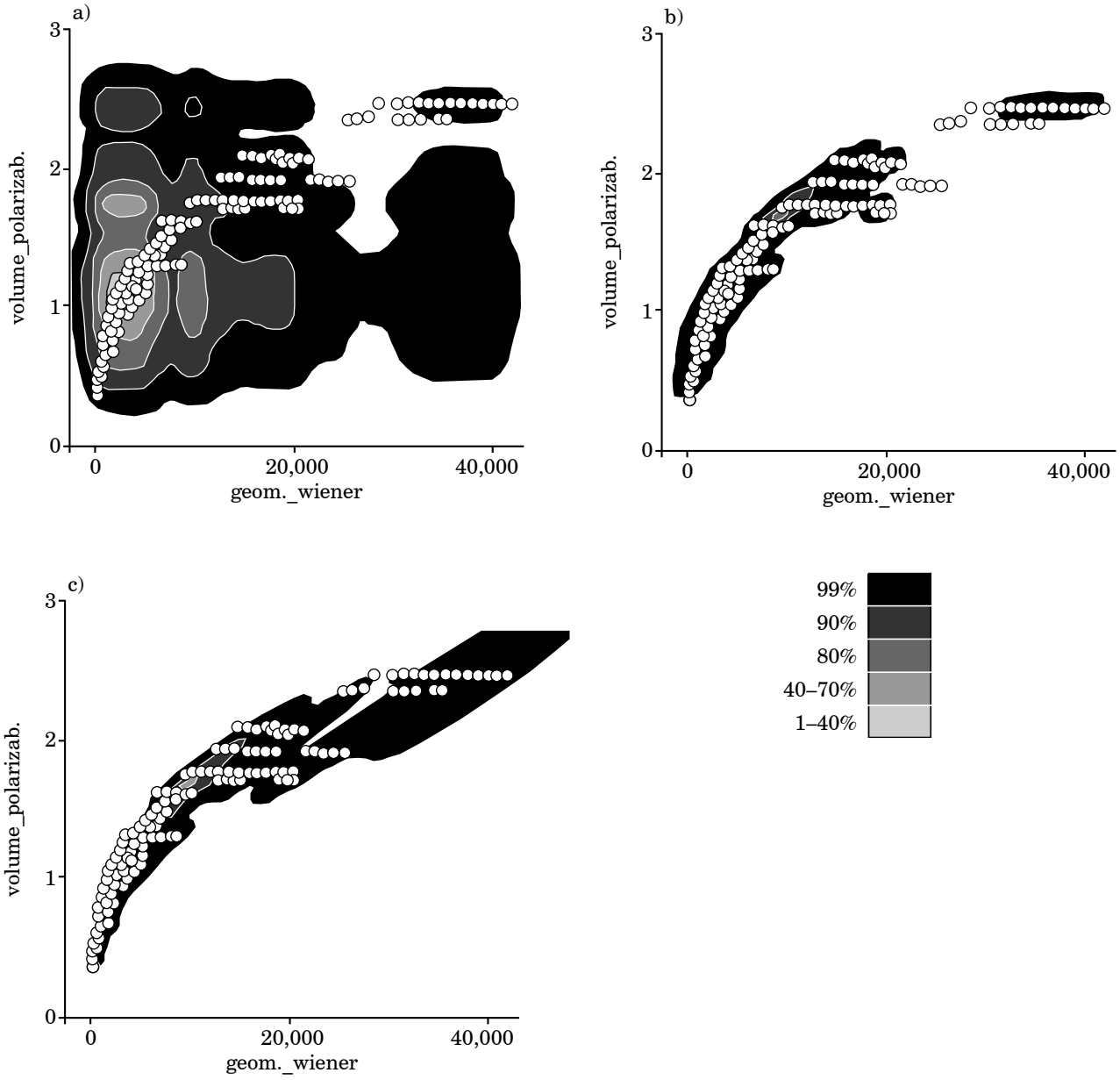
A HDR region has two main properties: 1) the density for every point inside the region is greater than the density for every point outside the region; and 2) for a given probability content, $(1 - \alpha)$, the interval is of the shortest length. It is not a trivial task to calculate the HDR, because it becomes computationally intensive unless one assumes a Gaussian or other parametric distribution (25). HDRs provide a very easy and intuitive interpretation: a point x lies in the region where $(1 - \alpha)$ points are situated or it has $(1 - \alpha)$ probability to belong to a set. This method overcomes the need to define *a priori* a cut-off value and reference point in a set, as required in distance-based methods.

The biggest challenge in HDR location is estimating the integral:

$$\int_{\substack{\text{xÂ,}\\ A=\{x:p(x)\le d\}}} p(x)dx = (1 - \alpha)$$

Applying an elementary integration algorithm in the multidimensional and non-parametric case results in a very high computational time. Nina Nikolova developed a novel, generic, fast method for HDR calculation for the non-parametric method. The novel algorithm was inspired by the basic idea of Monte Carlo integration — generate random points, evaluate function values at each point, calculate the sum of the values, and finally, multiply the sum by the multidimensional volume. The basic theorem of Monte Carlo integration (28, 29) estimates the integral of a function f over the multidimensional volume, where the angle brackets denote

Figure 6: 2D kernel probability density



(a) a product of 1D marginal densities in the original descriptor space, (b) a joint 2D kernel density estimation in the principal components space, (c) same as (b) with skewness correction.
X% represents x% High Density Region

taking the arithmetic mean over the N sample points:

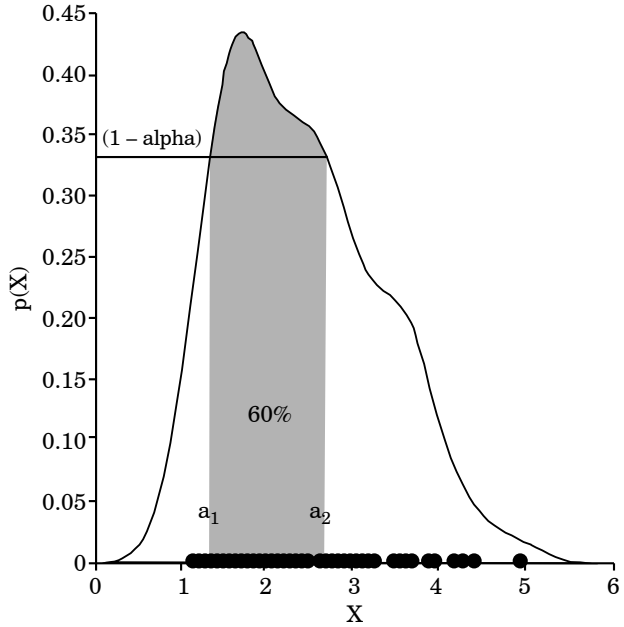
$$\int_{x \in S} f(x)dx \approx V\,[f(x)] \pm V \sqrt{\frac{(f^2) - (f)^2}{N}} \qquad (3)$$

$$[f] = \frac{1}{N}\sum_{i=1}^{N} f(x_i);\ [f^2] = \frac{1}{N}\sum_{i=1}^{N} f^2(x_i)$$

While the basic idea is simple, the algorithms for Monte Carlo integration of general functions are quite complex. However, we have the rather specific case of a probability density function, and it is possible to develop a simple, yet effective, algorithm. The algorithm consists of:

1. Setting the α value. (For example, if we are interested in regions covered by 90% of all data points, set $\alpha = 0.1$);

Figure 7: Probability density p(x) and 60% high density region



2. Evaluating $p(x_i)$ for each point x_i;

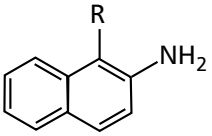
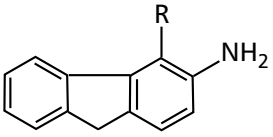
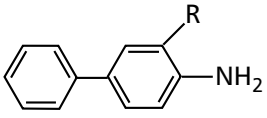
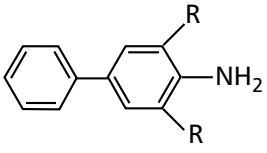
3. Sorting the points by descending $p(x_i)$ value; and

4. Counting the first M points belonging to the $(1 - \alpha)$ fraction of all the points.

The smallest p(x) of this $(1 - \alpha)$ set is the threshold value $d = D(1 - \alpha)$. For a query point y, p(y) is calculated. If $p(y) \geq D(1 - \alpha)$, then the point is within the dense region comprising $(1 - \alpha)$ of all the probability mass.

The density value, $d = D(1 - \alpha)$, is sufficient to assess whether a query point y will fall within the $(1 - \alpha)$ HDR. That is, we can determine whether a new compound is inside or outside the descriptor space covered by a given data set. The thresholds $D(1 - \alpha)$ for different α can be stored and used for further evaluations of query points. The knowledge of the density value is also sufficient for HDR visualisation (27, 30).

Structures of the Glende *et al.* (2001) test set[a]

Structure	R	No.	Compound	Abbreviation
	H	1	2-Aminonaphtalene	2-AN
	Et	2	1-Ethyl 2-aminonaphtalene	1-Et-2AN
	iPr	3	1-iPropyl 2-aminonaphtalene	1-iPr-2AN
	nBu	4	1-nButyl 2-aminonaphtalene	1-nBu-2AN
	tBu	5	1-tBu 2-aminonaphtalene	1-tBu-2AN
	H	6	2-Aminofluorene	2-AF
	Et	7	1-Ethyl 2- aminofluorene	1-Et-2 AF
	iPr	8	1-iPropyl 2- aminofluorene	1-iPr-2 AF
	nBu	9	1-nButyl 2- aminofluorene	1-nBu-2 AF
	tBu	10	1-tBu 2- aminofluorene	1-tBu-2 AF
	H	11	4-Aminobiphenyl	4-ABP
	Et	12	3-Ethyl-4-aminobiphenyl	3-Et-4 ABP
	iPr	13	3-iPropyl-4-aminobiphenyl	3-iPr-4 ABP
	nBu	14	1-nButyl 2- aminobiphenyl	3-nBu-4 ABP
	tBu	15	1-tBu 2- aminobiphenyl	3-tBu-4 ABP
	Me	16	3,5-Dimethyl-4-aminobiphenyl	3.5-diMe-4ABP
	Et	17	3,5-Diethyl-4-aminobiphenyl	3.5-diEt-4-ABP
	iPr	18	3,5-DiPropyl-4-aminobiphenyl	3.5-diiPr-4-ABP

[a]*The names of chemicals are not the IUPAC standard names, but have been retained as given in original paper (8).*