CADASTER

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Grant agreement no.: 212668

Collaborative project

Sub-Priority ENV2007 3.3.1.1: In-silico techniques for hazard-, safety-, and environmental risk-assessment

Work package 4: Integration of QSARs within hazard and risk assessment

Guidance on using QSAR models for probabilistic risk assessment

(Deliverable 4.2 - Report)

Due date of deliverable: 31 December 2012

Actual submission date: 31 December 2012

Start date of project: 1 January 2009 Duration: 4 years

Lead Contractor: National Institute of Public Health and the Environment (RIVM), Laboratory for Ecological Risk Assessment

Corresponding authors of document: Ullrika Sahlin[1] and Willie Peijnenburg[2]

1. Faculty of Science and Engineering, School of Natural Sciences, Linnaeus University, SE- 391 82 Kalmar, Sweden (ullrika.sahlin@lnu.se)

2. RIVM, Laboratory for Ecological Risk Assessment, PO Box 1, 3720 BA, Bilthoven, The Netherlands (willie.peijnenburg@rivm.nl)

Deliverable no. 4.2 – Guidance on using QSAR models for probabilistic risk assessment (report)

Project co-funded by the EU Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	x
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

General

CADASTER is a project that was granted within the 7th Research Framework Programme of DG Research of the European Commission. CADASTER aims at providing the practical guidance to integrated risk assessment within REACH by carrying out a full hazard and risk assessment for chemicals belonging to four compound classes. The main goal is to exemplify the integration of information, models and strategies for carrying out safety, hazard and risk assessments for a selected number of compounds within four specific chemical domains. Real hazard estimates will be delivered according to the basic philosophy of REACH of minimizing animal testing, costs, and time. CADASTER will show how to increase the use of non-testing information for regulatory decision whilst meeting the main challenge of quantifying and reducing uncertainty.

CADASTER has officially started on the 1st of January, 2009. The project officer on behalf of DG Research of the European Commission is Dr. Georges Deschamps, the project is coordinated by Dr. Willie Peijnenburg (RIVM).

Guidance on using QSAR models for probabilistic risk assessment

Summary

This document is deliverable 4.2. in the CADASTER project. It provides an overview of current guidance for the use of QSARs in risk assessment under REACH, and what has been done within the CADASTER project to address gaps in current guidance to facilitate the integration of QSARs in risk assessment.

Table of Contents

1. Introduction

The REACH regulation advocates the use of non-animal testing methods, but guidance is needed on how these methods should be used. The procedures to use alternatives to animal testing include methods such as chemical and biological read-across, in vitro results, in vivo information on analogues, (Q)SARs, and exposure-based waiving. The concept of Intelligent Testing Strategies (ITS) for regulatory endpoints that allow for efficient risk assessment has been outlined to facilitate the assessments. The CADASTER projected was initiated by the need to translate the ITS concept into a workable, consensually acceptable, and scientifically sound strategy. The optimization of the ITS concept should also be applicable within the precautionary principles that are put central in REACH. Therefore a main challenge within CADASTER has been to demonstrate the use of non-testing information for regulatory decision whilst meeting the main challenge of quantifying and reducing uncertainty.

Industry is primarily made responsible for carrying out the risk assessments, and practical guidance is therefore needed on how to apply the elements of the newly derived testing strategies in a consistent manner. During the execution of the CADASTER project, guidance for the use of non-testing methods in the European regulatory context has gone through major improvement. There is still a need of distinct application criteria and guidance on how to rigorously address uncertainty. This guidance document presents the achievements on the application of QSARs in probabilistic risk assessment, explicitly taking account of the inherent uncertainties associated with individual QSAR predictions. On the basis of the results obtained, application criteria are presented to aid judgment related to the use of QSARs as non-testing information supporting Chemical Safety Assessment.

A goal with the CADASTER project has been to exemplify the integration of information, models and strategies for carrying out safety-, hazard- and risk assessments for large numbers of substances. Methods have been derived for assessments made on a chemical class level, making use of the chemical domain as defined by molecular descriptors. Operational procedures of the integration of QSARs into probabilistic risk assessment have been developed, tested, and disseminated to guide a transparent evaluation of hazard and risk of emerging chemicals, explicitly taking account of variability and uncertainty in data and in models.

The purpose with this deliverable is to provide guidance for the integration of QSARs into chemical safety assessment, and more specifically in probabilistic risk assessment for which all important sources of uncertainties are to be characterized and propagated in the assessment. The structure of this report is made to guide the reader to relevant sources of information for the use of QSARs in chemical safety assessment. First, existing guidance for the use of QSARs under REACH are listed together with sources of relevance for the evaluation and integration of QSARs in probabilistic risk assessment. Then, unsolved issues of relevance are identified and followed up by what has been done within the project to address or solve these issues. Details of solutions and evaluations are given in appendices. The final section provides recommendation on future use of QSARs in risk assessment.

2. Overview of existing guidance

2.1. Guidance documents for probabilistic risk assessment

A thorough guidance on information requirements and Chemical Safety Assessment is found on the website of the Environmental Chemicals Agency (ECHA) available at http://echa.europa.eu/support:

Chapter R 19 in the Guidance on information requirements and Chemical Safety Assessment contains a description of probabilistic risk assessment under REACH and introduces uncertainty analysis.

There are many articles and books on risk assessment out of which two are mentioned here:

Jager, T., T. G. Vermeire, et al. (2001). "Opportunities for a probabilistic risk assessment of chemicals in the European Union." Chemosphere **43**(2): 257-264. *This can be seen as an introduction to the exposure and effect paradigm for probabilistic risk assessment.*

Aven, T. (2010). "Some reflections on uncertainty analysis and management." Reliability Engineering & System Safety **95**(3): 195-201. *This is a general introduction to probabilistic risk assessment that discusses the interpretation of probability and uncertainty in a broader perspective*

2.2. Guidance documents on the use of QSARs in Chemical Safety Assessment under REACH

Regulation (EC) No. 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH). Commission of the European Communities, Brussels. *This is the original regulatory document on the the use of QSARs in risk assessment under REACH.*

The following standard to ensure the validity of the use of QSARs plays a major role:

OECD (2006). Report on the regulatory uses and applications in OECD member countries of (Q)SAR models in the assessment of new and existing chemicals. Environmental Chemicals Agency, Finland.

Guidance documents found at the website of the Environmental Chemicals Agency (ECHA): http://echa.europa.eu/support

ECHA (2008). R.6: QSARs and grouping of chemicals, Guidance on information requirements and chemical safety assessment. Environmental Chemicals Agency, Finland.

ECHA (2009). Practical guide 5: How to report (Q)SARs? Environmental Chemicals Agency, Finland.

ECHA (2010). Practical guide 2: How to report weight of evidence? Environmental Chemicals Agency, Finland.

Reporting formats provide guidance on what is asked for to have a valid model and a valid prediction:

Joint Research Centre (2008). QSAR model reporting format (version 1.2), Institute for health and consumer protection.

Joint Research Centre (2008). QSAR Prediction Reporting Format (QPRF) (version 1.1). Institute for Health and Consumer Protection.

Case-studies on the use of QSARs in risk assessment:

Joint Research Centre (2011). A Framework for assessing in silico Toxicity Predictions: Case Studies with selected Pesticides. Report from the European Commission's Joint Research Centre, Institute for Health and Consumer Protection. *This contain a checklist of 10 key questions that the risk assessor should go through when evaluating a QSAR model in a regulatory purpose*

Pavan, M. and A. Worth (2008). "A set of case studies to illustrate the applicability of DART (Decision Analysis by Ranking Techniques) in the ranking of chemicals." JRC Scientific and Technical Reports EUR 23481 EN - 2008. *Refers to the use of QSAR predictions in classifications.*

The use of QSARs in risk assessment is exemplified and developed in several deliverables from research projects such as CADASTER (www.cadaster.eu), OSIRIS (http://www.osiris-project.eu/), EUFRAM (http://www.ist-world.org/ProjectDetails.aspx?ProjectId=641be1ad0b8244bdbecc4d8f56c1e068&SourceDatabaseId=081fd37e0ca64283be207ba37bb8559e) and OPENTOX (http://www.opentox.org/)

Two articles about the use of QSARs and their validity and reliability:

Eriksson, L., J. Jaworska, et al. (2003). "Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs." Environmental Health Perspectives **111**(10): 1361-1375. *The Bayesian framework for classification models is described in this paper.*

Gramatica, P. (2007). "Principles of QSAR models validation: internal and external." QSAR & Combinatorial Science **26**(5): 694-701.

Present Bayesian principles of inference and QSAR modelling

Aldenberg, T. (2004). "Review of methods for assessing the applicability domains of SARs and QSARs. Paper 3: Joint applicability domain and predictive uncertainty in QSAR regression."

Useful references for the judgement of confidence in a QSAR prediction:

Jaworska, J. and N Nikolova-Jeliazkova (2007). "How can structural similarity analysis help in category formation." SAR and QSAR in Environmental Research, **18**(3-4)

Jaworska, J., N. Nikolova-Jeliazkova and T Aldenberg (2005) "QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review." ATLA, **33**, 445–459.

Nikolova-Jeliazkova, N. and J. Jaworska (2005). "An Approach to Determining Applicability Domains for QSAR Group Contribution Models: An Analysis of SRC KOWWIN." ATLA, **33**, 461–470

T.I. Netzeva, A.P. Worth,T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W. Tong, G. Veith & C. Yang (2005). "ECVAM WORKSHOP REPORT Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships ." ATLA, **33**, 155-173

The AmbitDiscovery is a program that implement most metrics mentioned above and are used to evaluate predictive reliability for which download and manual is available at:

http://ambit.sourceforge.net/download_ambitdiscovery.html

3. Gaps in current guidance on the application of QSARs in probabilistic risk assessment addressed in the CADASTER project

Current guidance on the use of QSARs in chemical safety assessment does not provide enough support for the application of QSARs in probabilistic risk assessment. Identified gaps or needs are identified below and followed up by how these have been addressed within the CADASTER project.

3.1. A motivation to actually consider uncertainty in QSAR predictions in chemical safety assessment

Gap: Even though pointed out as relevant, is uncertainty in QSAR predictions given low or vague consideration in the available guidance or tools generating QSAR predictions.

CADASTER action: **Motivate why to consider uncertainty in QSAR predictions in chemical safety assessment**

Summary of the manuscript "Arguments for considering QSAR uncertainty in hazard and risk assessments" by Ullrika Sahlin, Laura Golsteijn, M. Sarfraz Iqbal, and Willie Peijnenburg in Appendix 1. The manuscript has been submitted to CADASTER workshop proceedings in ATLA.

3.2. Case-studies QSAR-integrated fate-, hazard- and risk assessment

Gap: More examples are needed to illustrate how QSAR predictions can be part of uncertainty analysis.

CADASTER action: **3.2. Perform Case-studies QSAR-integrated fate-, hazard- and risk assessment.**

Case-studies are presented in CADASTER deliverable 4.6. "Synthesis of research findings and recommendations for prioritization", out of which some are found in Appendix 2.

3.3. A conceptual framework that identifies uncertainty in QSAR prediction as being both quantitative and qualitative

Gap: Uncertainty is difficult to understand and is influenced by judgment and context. In order to provide guidance on how to consider uncertainty associated to the use of QSARs in probabilistic risk assessment there is a need of a common understanding of uncertainty in QSAR predictions.

CADASTER action: **3.3. Develop a conceptual framework that identifies uncertainty in QSAR prediction as being both quantitative and qualitative.**

Summary and table of the framework from the manuscript "Uncertainty in QSAR predictions" by Ullrika Sahlin in Appendix 3. The manuscript has been submitted to CADASTER workshop proceedings in ATLA.

3.4. An overview of approaches to quantify uncertainty in QSAR regression by probabilities

Gap: QSAR modeling covers a wide variety of supervised learning algorithms, which more or less quantify uncertainty in the error associated to individual QSAR predictions. There is need to describe and evaluate approaches to characterize the error in an individual QSAR prediction by a probability distribution.

CADASTER action: **3.4. Provide an overview of approaches to quantify uncertainty in QSAR regression by probabilities to stimulate progress on method development.**

Adopting the Bayesian framework for predictive inference may stimulate the integration of QSARs in probabilistic risk assessment, since it acknowledges both the use of expert judgment to build models and to quantify uncertainty in predictions. Since Bayesian principles in relation to QSARs most often are presented for classification models, activities within the CADASTER project have been focusing on applications on regression models. More details are found in Appendix 4.

3.5. Explore the possibilities of considering sources of variability and uncertainty in QSAR modeling

Gap: QSAR data are commonly given as point values for every compound, but these can be of varying quality and associated with different uncertainty or variability.

CADASTER action: 3.5. **Explore the possibilities of considering sources of variability and uncertainty in QSAR modeling**.

Adopting Bayesian modeling see Appendix 4 is applicable here. The benefit from consideration of variability in experimental values has been evaluated (Appendix 5.1 and 5.2).

3.6. Sensitivity analysis to evaluate the influence of qualitative QSAR uncertainty on the total confidence in an assessment

Gap: Qualitative uncertainty depends on to what extent a prediction is an extrapolation for a QSARs domain of applicability. There is a need to develop practical experience to support the judgment of confidence in assessment output that depends on qualitative uncertainty in QSAR predictions used as input parameters. When QSAR are used as input to an assessment, the judgment of whether a QSAR prediction is acceptable is made before the assessment is made. Judging what is acceptable or not may be sensitive to the context in which the assessment supports decisions.

CADASTER action: **3.6. Suggest a method for sensitivity analysis to evaluate the influence of qualitative QSAR uncertainty on the total confidence in an assessment.**

A slideshow presentation on info-quality analysis is included in Appendix 6 and an extended uncertainty analysis is described in the case-studies presented in Appendix 2. The communication of qualitative uncertainty has been stimulated by development of graphs of measure of predictive reliability versus prediction, together with the position of the training data set and the prediction.

3.7. QSAR-integrated SSD modeling

Gap: The added error from using a QSAR prediction compared to the experimentally based estimate ought to be considered when combining these kinds of information. We seek an approach to consider uncertainty in QSAR predictions that do not reduce variability in Species Sensitivity Distributions (SSD).

CADASTER action: **3.7. Evaluate the possibilities of QSAR-integrated SSD modeling**

The results of a simulation experiment demonstrating the influence of QSAR uncertainty on hazardous concentration assessed by the SSD approach, is given in Appendix 7.

4. Conclusions

QSARs are advantageous, compared to other non-testing methods, since their mathematical formulation open up for uncertainty to be quantified by modelling. A QSAR prediction is non-testing information that may replace unavailable testing information in Chemical Safety Assessment. It is not possible to say with 100% certainty that a QSAR prediction result in the same value as an estimate based on the corresponding experimental test. There is always and unknown error associated to the use of a predictive model instead of direct empirical observations. This error exist even if the relation between structures and activities (or properties) in a QSAR is strong. Also, experimental values can be erroneous, and that is why validity of data always is established before using a QSAR. Predictions from a QSAR are not better than the data used to build the model.

A successful communication of the uncertainty associated to QSAR predictions rely on a proper acknowledgement of error in a QSAR prediction in relation to the testing information it is meant to replace. This means that a QSAR prediction used in probabilistic risk assessment (i.e. where relevant sources of uncertainty are considered) ought to be foregone by consideration of the qualitative and quantitative characteristics of this error, which is specific for every individual prediction. A proper acknowledgement of uncertainty in QSAR predictions means to present predictions with their associated error either as an interval or as a probability distribution (i.e. quantitative characteristic), and a judgment of the confidence in the prediction given the QSAR models domain of applicability (i.e. qualitative characteristic). For this purpose methods to evaluate the validity and reliability of a QSAR may need to be complemented by assessment of uncertainty in an individual QSAR prediction and its influence on the risk assessment. The quantitative nature of QSARs based on chemical knowledge in combination with predictive inference, makes this into the non-testing method with the good possibilities of data-driven assessment of uncertainty as opposed to expert judgment only.

Uncertainty in risk assessment is usually described by probabilities, and is subjective reflecting the uncertainty of the risk assessor. A transparent characterization of uncertainty in QSAR predictions is informed by the underlying QSAR data, i.e. quantitative molecular descriptors and quantitative measures of an activity or property for a set of chemical compounds for which a QSAR is believed to exist, and inference from a probability model. Inference can be more or less dependent on the underlying QSAR algorithm, more or less based on expert judgment, but requires probabilistic modelling in addition to or integrated into the QSAR modelling. There is much to be gained by using probabilistic specifications of QSARs to allow for a quantification of uncertainty through inference on a statistical principle. The Bayesian principle of inference is valid given the usually small sizes of QSAR data and the way QSAR data are selected, in combination with the prevailing paradigm for understanding uncertainty in a risk context.

Based on the activities in CADASTER aimed to integrate QSARs into probabilistic risk assessment we propose the following recommendations:

General recommendations

- Support the use of QSAR predictions in risk assessment under REACH, but propose rigorous tools (such as uncertainty characterization and external validation) to avoid any misuse.
- Support the use of QSAR predictions by raising its advantages in relation to other in-Silico techniques.

Recommendations with respect to the development of future QSARs to facilitate the integration to risk assessment

- Build QSARs that are probabilistic such as Bayesian modelling.
- Build QSARs that model sources of variation in hierarchical levels to open up for the consideration of variability in experimental data, amongst others due to varying quality in underlying experimental data.
- Build QSAR integrated assessment models that in a hierarchical fashion integrate QSAR data, available experimental data, sources of variability and uncertainty to properly assess the hazard and risk endpoints.

Recommendations with respect to reporting of QSAR information

Extend/modify information requirements and reporting formats

- Such that uncertainty becomes naturally associated to a prediction
- To be able to consider uncertainty quantified by probabilities (e.g. it should be possible to attach a random sample from a predictive distribution to open up for Monte Carlo simulation).
- To include the assessment of quantitative uncertainty including evaluation of the approach taken (its theoretical bases and if possible by evaluation on QSAR data),
- To include a QSAR-specific recommended approach to judge the confidence in QSAR predictions (based on some metric to evaluate predictive reliability and reference cut-offs to aid judgement).

Recommendations with respect to the practical integration of QSARs into risk assessment

- There is a need for methods to propagate both qualitative and quantitative uncertainty associated to QSAR-predicted input parameters to risk assessment.
- There is a need to find simple rules of thumb that can be used to facilitate the reporting of uncertainty in QSAR predictions.

Recommendations for future activities

- Support workshops on the assessment and consideration of QSAR uncertainty with practical training.
- Support the development of more case-studies to show the impact and usefulness of considering uncertainty in QSAR predictions in the regulatory decision context under REACH.
- Support more research on the QSAR integrated assessments and development of user-friendly tools for uncertainty analysis and evaluation of quality assessment output with respect to quality in available background knowledge.

Appendices

Appendix 1

Title: Arguments for considering QSAR uncertainty in hazard and risk assessments

Ullrika Sahlin[1*], Laura Golsteijn[2], M. Sarfraz Iqbal[1], Willie Peijnenburg[3,4]

1. Linnaeus University, School of Natural Sciences, SE- 391 82 Kalmar, Sweden

2. Radboud University Nijmegen, Institute for Water and Wetland Research, Department of Environmental Science, PO Box 9010, 6500 GL, Nijmegen, The Netherlands

3. RIVM, Laboratory for Ecological Risk Assessment, PO Box 1, 3720 BA, Bilthoven, The Netherlands

4. Institute of Environmental Sciences, Leiden University, Leiden, The Netherlands

*corresponding author: Ullrika.Sahlin@lnu.se, fax: +46 480 44 73 40

Summary

Chemical regulation allows non-testing information to replace experimental values in hazard and risk assessments. Non-testing information on chemical activities or properties is subject to added uncertainty as compared to testing information, but this uncertainty is commonly not (fully) taken into account. Considering uncertainty in predictions from Quantitative Structure Activity Relationships (QSARs), a non-testing information, may improve the way QSARs support Chemical Safety Assessment under REACH. We argue that it is useful to consider uncertainty in QSAR predictions as it 1) supports rational decision making, 2) facilitates cautious risk management, 3) informs uncertainty analysis in probabilistic risk assessment, 4) may aid the evaluation of QSAR predictions in weight-of-evidence approaches, and 5) provides a probabilistic model to verify experimental data used in risk assessments. The discussion is illustrated by case-studies of QSAR integrated hazard and risk assessment from the EU-financed project CADASTER.

Key words: decision making, uncertainty analysis, probabilistic risk assessment, non-testing information

Manuscript is submitted to the CADASTER workshop proceedings in ATLA

Appendix 2

Case-studies conducted within CADASTER Work package 4: Integration of QSARs within hazard and risk assessment

Introduction to case-studies

An aim with CADASTER has been to provide practical guidance to QSAR-integrated risk assessment, by exemplifying the integration of information, models and strategies for carrying out safety-, hazard- and risk assessments for large numbers of substances. Some of the results are here presented as case-studies to demonstrate the use of QSARs (as an example of non-testing information) for regulatory decision whilst meeting the main challenge of quantifying and reducing uncertainty.

The purpose of the case-studies is to prioritize compounds by identifying those believed to be of highest concern based on information from QSARs. Such ranking based on QSAR predictions may be a fast way to scan over a large number of compounds. One could also rank based on the available information, i.e. compounds for which experimental data is lacking. The case-studies presented here demonstrate the integration of QSARs in chemical safety assessment and only use information retrieved from QSAR predictions, even though experimental data may be available for some compounds. The assessments and approaches have been selected to exemplify the considering uncertainty in QSAR predictions when integrating QSARs into hazard and risk assessment.

The case-studies presented here exemplify the use of in-silico structure-activity relationships (QSARs) and computational chemistry, including issues such as the applicability domain and validation status, and the use of probabilistic methods to consider variability and uncertainty. Uncertainty is here considered statistically by predictive inference, quantitatively by uncertainty analysis in a probabilistic assessment, and qualitatively through the evaluation and propagation of confidence in QSAR derived information. Variability is here considered through QSARs integrated to Species Sensitivities Distributions (SSDs).

Abbreviations

PFC Perfluoroalkylated substances and their transformation products, like perfluoroalkylated sulfonamides, alkanoic acids, sulfonates. Fluorinated compounds are typically a class of persistent, relatively hydrophilic compounds that may be toxic for man and environment.

BDE Polybrominated diphenylethers (PBDE), typically being a class of hydrophobic chemicals that pose a threat to man and the environment.

BTAZ Triazoles/benzotriazoles, a class of chemicals that are increasingly used as pesticides and anti-corrosives.

List of case-studies

Chemical group	Title	Corresponding author and affiliation	Reference
PBDEs	QSAR integrated fate assessment of PBDEs	S Iqbal LNU	D4.1 and manuscript in review(Iqbal, Golsteijn et al. in review)
PBDEs	QSAR integrated hazard assessment of PBDEs	E Rorije RIVM	D4.1
PBDEs	Uncertainty analysis in QSAR integrated hazard assessment of PBDEs	U Sahlin LNU	Appendix PBDE.1
PBDEs	Prioritization based on PBT evaluation of PBDEs	U Sahlin LNU	Appendix PBDE.1
PBDEs	Impact assessment of PBDEs	A Shipper RUN	D4.4 and two manuscripts (details in appendix PBDE.2)
Triazoles	QSAR integrated fate and effect assessment of triazoles	L Golsteijn RUN	D4.1 and manuscript in review (Golsteijn, Iqbal et al. in review)
Triazoles	Prioritization based on hazard assessment of BTAZs	U Sahlin LNU	Appendix BTAZ.1
Triazoles	Prioritization based on risk assessment of BTAZs	U Sahlin LNU	Appendix BTAZ.2
Triazoles	Uncertainty analysis of QSAR integrated hazard assessment of BTAZs	U Sahlin LNU	Appendix BTAZ.3
Triazoles	Uncertainty analysis of QSAR integrated risk assessment of BTAZs	U Sahlin LNU	Appendix BTAZ.4
PFCs	QSAR integrated hazard assessment of PFCs	U Sahlin LNU	Appendix PFC.1
PFCs	Prioritization based on hazard assessment of PFCs	U Sahlin LNU	Appendix PFC.1
PFCs	QSAR integrated fate assessment of PFCs	L Golsteijn RUN	Appendix PFC.2 (Golsteijn, Papa et al. manuscript)

Blevins, Laura Golsteijn, Mark AJ Huijbregts, M Sarfraz Iqbal, Willie Peijnenburg, Emiel Rorije and Igor Tetko (2012). "Application of QSAR models for probabilistic risk assessment (report and model)." Available at

http://www.cadaster.eu/sites/cadaster.eu/files/data/deliverable/public/Deliverable4.1.pdf

Manuscripts submitted or in preparation

Golsteijn, L., M. S. Iqbal, et al. (in review). "The Relative Importance of Uncertainty in Predicted Chemical Properties for the Comparative Toxicity Potentials of Triazoles ".

Golsteijn, L., E. Papa, et al. (manuscript). "The Role of Uncertain Koc Predictions in the Overall Persistency and Long Range Transport Potential of Perfluorinated Chemicals."

Iqbal, M. S., L. Golsteijn, et al. (in review). "Dealing with QSPR predictive uncertainty in environmental fate modeling ".

Appendix 3
Title: Uncertainty in QSAR predictions

Ullrika Sahlin

School of Natural Sciences, Linneaus University, Kalmar, Sweden

Ullrika.Sahlin@lnu.se

Fax: +46 480 44 73 40

Summary

It is relevant to consider uncertainty in individual predictions when Quantitative Structure - Activity or Property Relationships (QSARs) are used to support decisions of high societal concern. Successful communication of uncertainty in the integration of QSARs in Chemical Safety Assessment under REACH can be facilitated by a common understanding of how to define, characterize, assess and evaluate uncertainty in QSAR predictions. A QSAR prediction is, compared to experimental estimates, subject to added uncertainty that comes from using a model instead of empirically based estimates. A framework is provided that distinguish between uncertainty in a QSAR prediction being quantitative, i.e. for regressions related to the error in a prediction and characterized by a predictive distribution, and qualitative, expressing our confidence in using the model to predict a particular compound judged based on a quantitative measure of predictive reliability. A quantitative (i.e. probabilistic) predictive distribution is possible to assess given the supervised learning algorithm, the underlying QSAR data, a probability model for uncertainty and a statistical principle for inference. The integration of QSARs into risk assessment may be facilitated by including the assessment of predictive error and predictive reliability into the “unambiguous algorithm” as asked for by the second OECD principle.

Key words: regression, knowledge-based uncertainty, probabilistic risk assessment, uncertainty

analysis, applicability domain

Submitted to CADASTER workshop proceedings

Tables

Table 1. A framework for the definition of a QSAR prediction, which dependent on the purpose of prediction can be without or without uncertainty.

ID	Description	Notation
1	Quantitative descriptor(s)	X [a]
2	Quantitative measure of a property or activity	Y
3	QSAR	Y | X [b]
4	Known values on Y	y
5	Specific values for the i'th compound	$\{y,X\}_i$
6	QSAR data is a set of n compounds for which the quantitative property or activity is known	$\{y,X\}_{i = 1:n}$
7	Supervised learning algorithm	A
8	QSAR model is a supervised learning algorithm and QSAR data	Y | X, $\{y,X\}_{i=1:n}$, A
9	Property or activity of query compound	Z
10	Known values on Z	z
11	Quantitative descriptor(s) of query compound	W [a]
12	QSAR external data is a set of nExt compounds with known values but not used to train the model	$\{z,W\}_{j = 1:nExt}$
13	QSAR prediction without uncertainty is a supervised learning algorithm, QSAR data and descriptors for the query compound	Z | W, $\{y,X\}_{i=1:n}$, A
14	Algorithm to assess uncertainty	U
15	QSAR prediction with uncertainty is a supervised learning algorithm that includes the assessment of uncertainty, QSAR data (sometimes including external QSAR data) and descriptors for the query compound	a) Z | W, $\{y,X\}_{i=1:n}$, AU or b) Z | W, $\{y,X\}_{i=1:n}$, $\{z,W\}_{j=1:nExt}$, AU

[a] We assume the values on X and W always are known.

[b] The symbol "|" stands for "given that"

Appendix 4

An overview of approaches to quantify uncertainty in QSAR regression by probabilities

Text is modified from "Uncertainty in QSAR predictions" submitted to CADASTER workshop proceedings in ATLA by Ullrika Sahlin

A prerequisite for successful communication of uncertainty is to understand what is meant by a QSAR prediction and its uncertainty. QSARs can roughly be divided into two kinds of predictions: classifications and regressions (Eriksson, Jaworska et al. 2003). A classification places a compound in one out of at least two categories, such as biodegradable or not. Uncertainty assessments for classifications may be based on contingency table statistics (Fielding and Bell 1997) and the assessment of the probability of making a correct classification given descriptor values is done in a Bayesian framework (Eriksson, Jaworska et al. 2003). Regression within the QSAR community means modelling of a continuous response, such as boiling point. The assessment of uncertainty in QSAR regressions is hampered by the prevailing point prediction view on QSAR predictions, and difficult to grasp given the wide array of modelling approaches that more or less model uncertainty in predictions (Sahlin, Filipsson et al. 2011). As a start, there is a need for a common understanding in uncertainty and its assessment with a broad perspective on modelling algorithms.

The focus here is on the assessment of uncertainty given the way QSAR modelling commonly consider available experimental data, which is as experimentally based point estimates. Even though relevant, the uncertainty that comes from both variability and measurement errors in experimental data cannot be quantitatively assessed unless modelled and recognized in the QSAR data (but see Tebby and Mombelli 2012). Further, the focus on quantitative approaches, algorithms and metrics does not mean that we reduce the importance of the chemical perspective for a successful implementation of QSARs in chemical regulation.

The meaning of QSAR uncertainty needs to be understood in relation to risk assessment practice. Risk assessment is a science-based approach, but nevertheless the characterization of uncertainty rest upon assumptions and decisions taken by the risk assessor. In general one may view uncertainty in a risk assessment as reflecting the risk assessor's uncertainty in predicted quantities to express risk, given available background knowledge (Aven 2010). Thus uncertainty, in the context of risk assessment, is to be understood as a subjective judgment that can change when new data, models or expert knowledge are added to the background knowledge. In order to tally the interpretation of uncertainties in input with that of the output of an assessment, the final interpretation of uncertainty in input parameters supported by QSAR predictions before entering an assessment will be as a subjective judgement of scientific basis for decision support (National Research Council 2009). However, even though uncertainty with the purpose to support risk assessment should be assessed to reflect the risk assessor's uncertainty in a QSAR prediction, its assessment can more or less based on data and probabilistic modelling, and is not constrained by the QSAR algorithm. Even though uncertainty is subjective, there is a need for unambiguous algorithms for its assessment to inform and support the final choice of its characterization.

Uncertainty in QSAR predictions is a major concern, especially when these predictions may influence human and animal lives as well as the safety of environmental systems. Approaches to assess uncertainty must therefore be as correct as possible, which poses a need to evaluate the quality in assessments of uncertainty, of which transparency, repeatability, and motivation are relevant characteristics. To aid we seek a useful guidance to characterize uncertainty in a QSAR prediction with the purpose to support decision making pointing at its definition, characterization, assessment and evaluation (Table 1).

Table 1. The characterization of uncertainty in a QSAR prediction useful to support decision making requires an unambiguous definition, characterization, assessment and evaluation.

Uncertainty	Predictive error	Predictive reliability
Definition	Magnitude of the add error in a prediction compared to experimental based estimate	Confidence in using a model to predict a specific compound
Characteristic	Quantitative probability distribution	Qualitative judgment of confidence (e.g. high or low)
Assessment	Probabilistic modelling of the error based on sampling, re-sampling, or probability theory maybe in combination with expert judgment	Confidence assessed by expert judgment (informed by relative measures such as density, distance and variation in perturbed predictions) or empirical coverage
Evaluation	Empirical coverage for a chosen level of confidence or likelihoods for an external data set (relative)	Difficult to evaluate a qualitative judgment per se. Alternative measures of predictive reliability can be evaluated for their abilities to capture a trend in predictive error or perceived lower reliability.

Chemical Safety Assessment asks for knowledge-based uncertainty in QSAR prediction in the relation to information requirements and uncertainty analysis. Information must fulfil several requirements before allowed to support a Chemical Safety Assessment (ECHA 2008). In particular ECHA asks for the validity of a selected QSAR to have been assessed and that it must have been verified that the chemical predicted falls within the applicability domain of to give a reliable prediction (ECHA 2009). The last requirement is a qualitative characterization of uncertainty (which we refer to as predictive reliability) related to the use of a QSAR to predict a specific chemical.

Uncertainty analysis is conducted to evaluate the need to refine an assessment and to inform the risk assessor on the magnitude of risk. Uncertainty can be analysed in three tiers with increasing precision of the quantification of uncertainty, going from deterministic, worst (plausible case) to probabilistic. Probabilistic means that uncertainty is given a full characterization by specifying likelihoods of all possible values an input parameter may take quantified by probabilities. Uncertainty analysis distinguish between parameter uncertainty, model uncertainty and scenario uncertainty (ECHA 2008). Out of these three, uncertainty in a QSAR prediction is most closely associated to parameter uncertainty, which according to ECHA "is the uncertainty involved in the specification of numerical values". Parameter uncertainties include measurement errors, sample uncertainty, selection of the data used for assessing the risk, and extrapolation uncertainty, where latter can be "the use of alternative methods (e.g. QSAR, in-vitro test, read-across for similar substances) or use of assessment factors (e.g. inter-species, intra-species, acute to chronic, route to route, lab to field extrapolation)". The need to quantify uncertainty in QSAR predictions by probabilities was pointed out by Walker et al (2003) suggesting *"that errors needs to be evaluated when applying QSARs by providing confidence intervals that take into consideration the uncertainty associated with the estimate"*, and we note that a confidence interval presumes an underlying probability distribution.

This shows that there is a need to characterize uncertainty in a QSAR prediction that is both qualitative, expressing our confidence in using a prediction to support decision making, and quantitative, expressing our belief in what values the predicted property or activity may have after being observed (Sahlin, Filipsson et al. 2011). The qualitative uncertainty is related to reliability in individual predictions, and we refer to this as "predictive reliability" to avoid confusing it with reliability in a more general meaning (Table S1). The quantitative uncertainty is associated to the predictive error, which is measure describing the distance between a point prediction and the actual value, and may change from compound to compound. Both predictive reliability and predictive error are sensitive to which degree a prediction is an extrapolation from a model.

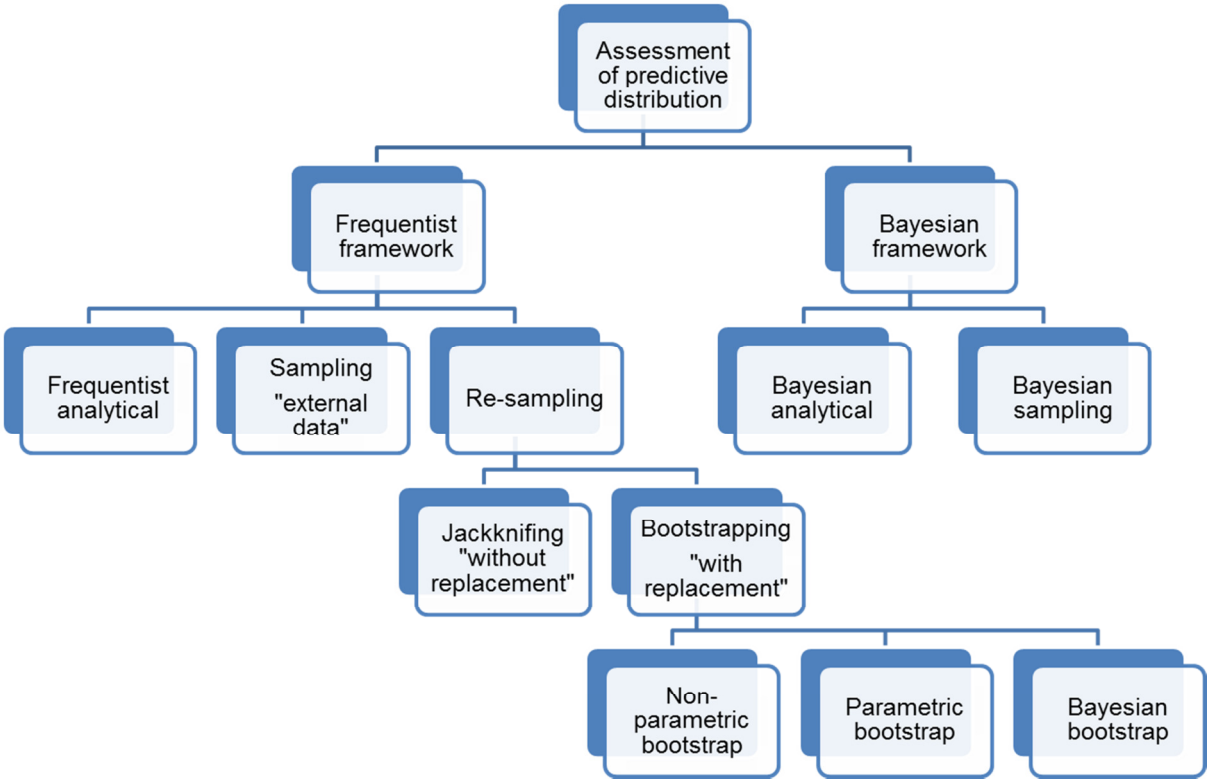


Figure 1. An overview of approaches to quantify uncertainty in QSAR regression by probabilities.

Characterizing quantitative QSAR uncertainty – the predictive error

Definition and characterization

Quantitative uncertainty is related to the error in a prediction which for a regression is the difference between the unobserved and predicted property or activity of a compound. We aim to quantify uncertainty in the associated error to an individual prediction by a probability distribution that express our belief in what values the error in a prediction may take after the property or activity has been observed. The probability distribution for the predictive error is here referred to as the predictive distribution.

Assessment

Approaches to assess the predictive distribution can be made under different statistical frameworks, using more or less data intensive methods, with more or less specified probabilistic models for uncertainty (Figure 1). Sampling Theory assess predictive error based on a representative sample. Such (frequentist) inference rests upon assumptions of independent and, for example, identically distributed observations, in combination with a probabilistic assumption of uncertainty. Under violence of any of these assumptions, appliers of frequentist inference run into problems. The Bayesian paradigm for inference *assign*, instead of *assume*, a probabilistic model for observations, and assign models for uncertainty in parameters (so called priors). Bayesian inference uses Bayes rule to update expert knowledge with information in empirical observations. The result is a well-defined

probabilistic model of uncertainty. In cases of doubts, the caveat is the necessity to choose priors and probabilistic models (likelihoods). For example, there is no need to check an assumption of normality of errors (as in the frequentist case), as this is assigned through expert judgment.

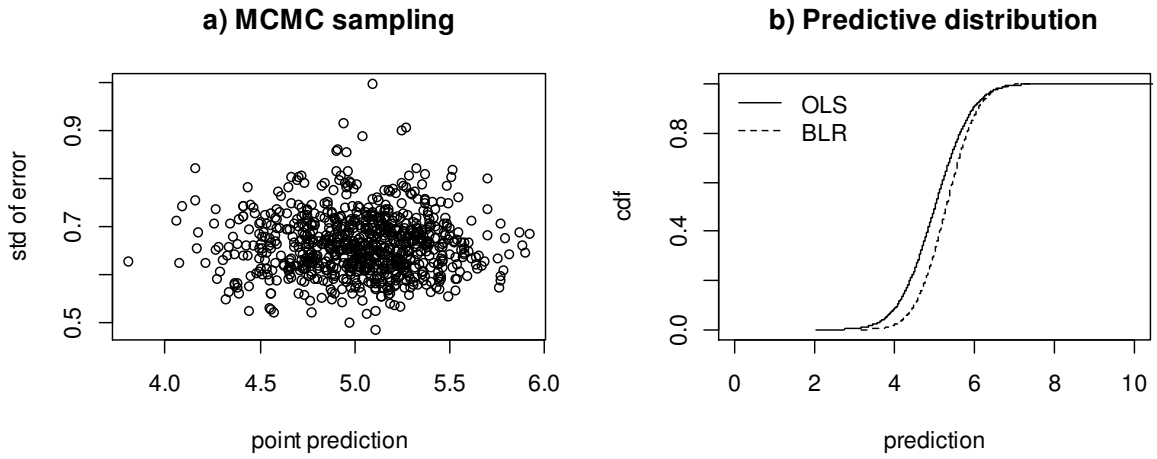


Figure 2. An illustration of a MCMC sample of a triazole predicted by the B(TAZ) QSAR (a) and the resulting predictive distribution which is compared to a predictive distribution based on an Ordinary Least Squares (OLS) regression fitted to the same descriptors (b). The predictive distribution for the OLS is assessed by assuming errors to be independent and identically distributed as Normal distribution with fixed variance, which generates a predictive distributions being a Student-t.

A third approach is to assign a probability distribution for predictive error based on expert judgment only. This can for example, be based on experience of experimental testing, or based on combinations of different sources of information. Sampling Theory and solid expert judgment can be seen as extremes kinds of Bayesian inference. Expert judgment of uncertainty can be seen as Bayesian modelling of the error with no updating, i.e. based on prior distribution only. The difference between frequentist and Bayesian inference in parametric and non-hierarchical linear regressions is usually negligible given a large data set or non-informative. Differences between predictive distributions generated by the Bayesian Lasso and Student-t following frequentist statistical inference from an OLS prediction (see e.g. Montgomery, Peck et al. 2001) in Figure 2, may derive from the use of informative priors in combination with a weak signal in the QSAR data. Note that the comparison between OLS and Bayesian is made for a given set of descriptors. Perhaps the largest difference between the two approaches lies in the selection of descriptors.

In order to do assess the predictive distribution a probability model is to be included or added to the supervised learning algorithm for prediction. There is a need to acknowledge and discuss the conditions and suitability of different approaches to include the assessment of predictive distribution in the QSAR algorithm.

Work in CADASTER project has been done to develop models/methods to assess predictive error, either as an absolute (point) value of the error or as the predictive distribution. Here follows a description of what has been done with suitable references to CADASTER outcomes.

Bayesian analytical

Result for Student – t are found in Deliverable 4.1 and in case-studies (Golsteijn, Iqbal et al. in review; Iqbal, Golsteijn et al. in review; Golsteijn, Papa et al. manuscript).

Bayesian sampling

The predictive distribution have been characterized by Markov Chain Monte Carlo (MCMC) sampling from a QSAR modelled by Bayesian lasso in the consensus modelling of aquatic toxicity for three species (Cassini, Kovarich et al. in review). The Bayesian lasso has been implemented as an alternative modelling algorithm on the CADASTER web tool.

Sampling

Assessment of predictive error as the average error evaluated on an external data set have been done for the majority of models developed in CADASTER (e.g. Tetko, Sushko et al. 2008; Papa, Kovarich et al. 2009)

Re-sampling without replacement

This approach is what is done when using any kind of cross-validation such as Leave-One-Out (LOO) or Leave-M-Out assessment of predictive error based on Predictive Error Sum of Squares. This approach has been applied on the majority of models developed in the project. It is also considered in on of CADASTER simulation studies (Sahlin, Jeliazkova et al. In review).

Re-sampling with replacement

Non-parametric bootstrap

A probabilistic characterization of the predictive error without any specific (i.e. parametric) probability distribution have been assessed based on modified residuals (Sahlin, Jeliazkova et al. In review). In this approach a sample from the empirical distribution of residuals, modified under a chosen set of assumptions, are used to assess predictive distribution.

Parametric bootstrap

Error from QSAR regressions can be assumed to have a symmetric predictive distribution such as Gaussian or Student-t. Predictive distributions were defined by local assessment of predictive error that specified the standard deviation in an assumed Gaussian distribution. Method development have been made on modified residuals (Sahlin, Jeliazkova et al. In review) and segment-based assessments (Tetko, Sushko et al. 2008; Zhu, Tropsha et al. 2008). Local assessment of characteristics of accuracy for classification models has med developed as well (Sushko, Novotarskyi et al. 2010).

Bayesian bootstrap

Bayesian bootstrap is a combination of Bayesian sampling and bootstrapping, where the sampling is made based on prior probabilities (Davison and Hinkley 1997).

Evaluation

In the same way as QSAR algorithm needs to be verified, ought also the algorithm to assess predictive error by the predictive distribution needs to be evaluated for a particular QSAR. Sampling and re-sampling are data rich methods and in the domain where there are many data points. As the size of samples become smaller it increases the need for parametric specification of QSAR model including the probabilistic model for uncertainty. Some modellers may feel uncomfortable with adding prior information or making parametric specifications. It is therefore important to evaluate a model, by for example comparing it to a less constrained alternative. The burden of information and resources for calculations should be in relation to the required level of detail of the quantified uncertainty (Jager, Vermeire et al. 2001; ECHA 2008). Before suggesting a resource demanding and complex approach to assess uncertainty, it may be relevant to evaluate it in relation to a simpler

assessment. A candidate for a Rule of Thumb, which can be read out from information provided in QMRF, is to assign a Gaussian distribution with the point prediction and a reported value on Mean Square Error of Prediction as its first and second moment (Sahlin, Filipsson et al. 2011).

Measures to evaluate predictions with uncertainty are empirical coverage, which should show a one-to-one correspondence to the assigned confidence levels (see examples in Figure 3). There are also relative measures based on likelihood-based measures that can be used to make relative comparison between alternative algorithms (see e.g. Tetko, Sushko et al. 2008) or as weights for model averaging of alternative predictions in consensus modelling (Johnson and Omland 2004).

References

Aven, T. (2010). "Some reflections on uncertainty analysis and management." Reliability Engineering & System Safety **95**(3): 195-201.

Cassini, S., S. Kovarich, et al. (in review). "Evaluation of CADASTER QSAR models for aquatic toxicity of (benzo-)triazoles and prioritization by consensus."

Davison, A. C. and D. V. Hinkley (1997). Bootstrap methods and their application. Cambridge, Cambridge Univ. Press.

ECHA (2008). R.6: QSARs and grouping of chemicals, Guidance on information requirements and chemical safety assessment.

ECHA (2008). R.19 Uncertainty analysis, Guidance on information requirements and chemical saftey assessment.

ECHA (2009). Practical guide 5: How to report (Q)SARs?, European Chemicals Agency Helsinki Finland.

Eriksson, L., J. Jaworska, et al. (2003). "Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs." Environmental Health Perspectives **111**(10): 1361-1375.

Fielding, A. H. and J. F. Bell (1997). "A review of methods for the assessment of prediction errors in conservation presence/absence models." Environmental Conservation **24**(1): 38-49.

Golsteijn, L., M. S. Iqbal, et al. (in review). "The Relative Importance of Uncertainty in Predicted Chemical Properties for the Comparative Toxicity Potentials of Triazoles ".

Golsteijn, L., E. Papa, et al. (manuscript). "The Role of Uncertain Koc Predictions in the Overall Persistency and Long Range Transport Potential of Perfluorinated Chemicals."

Iqbal, M. S., L. Golsteijn, et al. (in review). "Dealing with QSPR predictive uncertainty in environmental fate modeling ".

Jager, T., T. G. Vermeire, et al. (2001). "Opportunities for a probabilistic risk assessment of chemicals in the European Union." Chemosphere **43**(2): 257-264.

Johnson, J. B. and K. S. Omland (2004). "Model selection in ecology and evolution." Trends in Ecology & Evolution **19**(2): 101-108.

Montgomery, D. C., E. A. Peck, et al. (2001). Introduction to linear regression analysis. New York, Wiley.

National Research Council (2009). Science and decisions : advancing risk assessment. Washington, D.C., National Academies Press.

Papa, E., S. Kovarich, et al. (2009). "Development, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers." Qsar & Combinatorial Science **28**(8): 790-796.

Sahlin, U., M. Filipsson, et al. (2011). "A Risk Assessment Perspective of Current Practice in Characterizing Uncertainties in QSAR Regression Predictions." Molecular Informatics **30**(6-7): 551-564.

Sahlin, U., N. Jeliazkova, et al. (In review). "Applicability domain dependent predictive uncertainty in QSAR regressions."

Sushko, I., S. Novotarskyi, et al. (2010). "Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set." Journal of Chemical Information and Modeling **50**(12): 2094-2111.

Tebby, C. and E. Mombelli (2012). "A Kernel-Based Method for Assessing Uncertainty on Individual QSAR Predictions." Molecular Informatics **31**(10): 741-751.

Tetko, I. V., I. Sushko, et al. (2008). "Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection." Journal of Chemical Information and Modeling **48**(9): 1733-1746.

Walker, J. D., L. Carlsen, et al. (2003). "Improving opportunities for regulatory acceptance of QSARs: The importance of model domain, uncertainty, validity and predictability." Qsar & Combinatorial Science **22**(3): 346-350.

Zhu, H., A. Tropsha, et al. (2008). "Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis." Journal of Chemical Information and Modeling **48**(4): 766-784.

Appendix 5.1

Explore the possibilities of considering sources of variability and uncertainty in QSAR modeling

Varying quality in QSAR data is possible to integrate through Bayesian hierarchical modeling, where quality may be assigned by expert judgment (Willighagen et al. 2011). Experimental variability has been considered by re-sampling (e.g. Tebby and Mombelli 2012). Within CADASTER a simulation experiment have been performed on the added value of considering multiple experimental values compared to their averages as QSAR data. The conclusion is that considering multiple data by the method of weighted linear regression did not outperform building a model on average values. These results are reported in a Master thesis in Environmental Science at Linneaus University, Sweden and poster at second CADASTER workshop in Munich 2012.

Willighagen, E. L., J. Alvarsson, et al. (2011). "Linking the Resource Description Framework to cheminformatics and proteochemometrics." J Biomed Semantics **2 Suppl 1**: S6.

Tebby, C. and E. Mombelli (2012). "A Kernel-Based Method for Assessing Uncertainty on Individual QSAR Predictions." Molecular Informatics **31**(10): 741-751.

Predictive uncertainty may be improved by efficient use of experimental information for QSARs

- Weighting versus averaging in linear regression

Marit Strömberg and Ullrika Sahlin*

Linnaeus University, School of Natural Sciences, Kalmar/Växjö, Sweden.

*corresponding author

Introduction and Aim

The use of QSARs in chemical regulation or other applications of decision making is possible if they provide predictions with acceptable confidence (Comber et al. 2003, Cronin et al. 2003). Confidence is evaluated in terms of a model's predictive ability, which includes a precise assessment of uncertainty in predictions. Uncertainty in predictions from QSAR models arises not only from the strength of the analogy assumption, saying that molecules with similar structure should have similar physiochemical properties, but also from the application of a statistical model/learning algorithm and from the quality of the experimental data (Schultz et al. 2003, Trophsa 2010). Experimental data show variation e.g. from having experiments done at different labs. Even though there may be more than one experimental value for a given compound, QSARs are today mostly developed by using only one experimental value for each compound, selected by expert judgment or as averages (Papa et al. 2009). The question that arises is if consideration of more information in empirical data in QSAR-development may improve the predictive ability of the model. There are examples where differences in quality of measurement methods have been considered by weighting motivated as prior information based on expert judgment (Willighagen et al. 2011).

The **aim** was to compare predictive ability of QSARs developed on several experimental values per compound to QSARs developed on averaged experimental values.

Table 1. QSARs in Papa et al. (2009)

	Model ID			
Characteristic	2	3	5	6
Endpoint	T_M	Log(1/P_i)	LogK_{oa}	LogK_{ow}
Descriptor	X2A	T(O..Br)	T(O..Br)	T(O..Br)
Number of training data	20	28	24	14
Number of multiple measurements	7	6	5	6
Number of test data	5	6	6	6

Table 3. Comparison of LRW and LRAV based on Papa et al.'s (2009) models.

	Model ID			
Statistic	2	3	5	6
difference in log likelihood score	0.08	0.02	0.07	0.80
Kolomogorov Smirnov Stat. (LRAV)	0.51	0.37	0.31	0.37
Kolomogorov Smirnov Stat. (LRW)	0.40	0.37	0.26	0.23

Fig 1

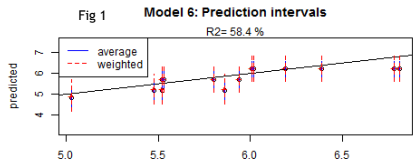
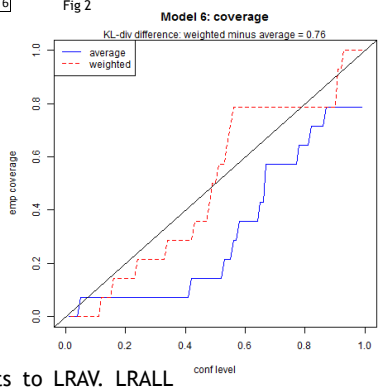


Fig 2



Models and Analysis

Multiple point estimates was considered by building weighted linear regressions with weights assigned such that each chemical had equal contribution to the loss function in the least squares regression. The weighted linear regression (LRW) and the linear regression based on all experimental data (LRALL) were each compared to the linear regression based on averages (LRAV). The modeling approaches ability to predict (including to assess predictive uncertainty) were evaluated by

1) The correlation between predicted and observed values in an external test data set,
2) Empirical coverage to theoretical confidence levels, and
3) Log likelihood scores derived for a common external data set under the corresponding predictive distributions.

Predictive uncertainty was here assessed as a non-parametric distribution by model-based bootstrap.

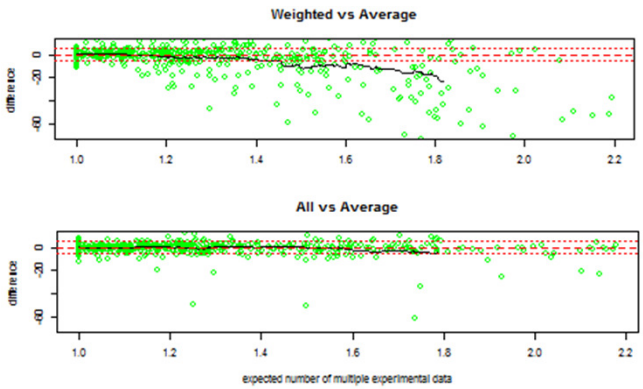
First, the effect of considering more experimental information was evaluated on four QSAR data sets from models developed by Papa et al. (2009) (Table 1). Second, in order to seek generality artificial datasets were constructed (Table 2). The comparisons were done on models judged as having good predictivity on average, which were those with R^2>0.6 for the training data, and where at least one of the approaches succeeded reasonably well in assessing the predictive uncertainty[1]. Differences in performance between modeling approaches were evaluated by the difference in logged likelihood scores, where a difference within 5 is "barely worth mentioning"[2].

[1] Judged as those with a Kolmogorov Smirnov statistic < 0.2, i.e. a significance level of 0.05
[2] According to the decibans scale for Bayes' factor.

Table 2. Characteristics of the artificial QSAR data sets

Model	Formula/Distribution
Endpoint values, Y =	X·B + random error + model error
Descriptor values, X	uniform(0,1)
Regression coefficients, B	uniform(0,1)
Random errors	normal(0,1) ·chisquare(1,s)
Model errors	normal(0,e)
Number of measurements per chemical	bionomial(m,p)
Characteristic	**Range**
Probability of multiple measurements, p	0-0.3
Size of multiple measurements, m	1-5
Size of training data, n	10-50
Random error, s	0.01-0.3
Model error, e	0.01-0.3
Number of descriptors, k	1-4

Fig 3. Differences in log likelihood score with a trend shown by moving average. The dotted lines indicate the zone where the differences between modeling approaches are "barely worth to mention". A negative difference favors using averages of experimental values.



Results

LRW rendered identical regression coefficients to LRAV. LRALL gave slightly different regression parameters (Fig 1). The estimates of model error and thereby uncertainty in predictions for the three models were all different (Fig 2). All of the four models by Papa et al. (2009) LRW showed an improved predictivity as compared to LRAV (Table 3) indicating that uncertainty in QSAR predictions may be improved by using weighting instead of averaging.

Regarding the QSAR models based on the artificial data sets none of the three modeling approaches had always better predictive performance than the others, and most differences in models' prediction ability were within the "barely worth mentioning" zone (Fig 3). LRW performed on average worse than LRAV, and the performance got worse with increasing expected number of experimental values per compound (p-value less than 0.001). LRALL had a slightly lowered performance, compared to LRAV, with increasing expected experimental values as well (p-value less than 0.01). Neither the number of compounds per descriptor nor expected total variance influenced the relative performances of the models.

Conclusion

The general conclusion is that of the three investigated model types there is no specific model type that always is in favor in terms of model predictivity, and which approach that is best depends on the specific data set. Therefore it could be worthwhile to consider all three types when developing a QSAR by linear regression.

Acknowledgement
This study was funded by the FP7 project CADASTER (grant agreement 212668). For more information, visit www.cadaster.eu

References
Comber et al (2003). Environ. Tox. and Chemistry, 22(8): 1822-1828; Cronin et al (2003). Environ. Health Persp. 111(10): 1376-1390; Papa et al (2009). QSAR & Comb Science. 28(8): 790-796; Schultz and Cronin (2003). Environ. Tox. and Chemistry. 22(3): 599-607; Trophsa (2010). Mol. Inf, 29: 476-488; Willighagen et al (2011). J. Biomedical Semantics. 2(Suppl 1):S6

Quality in information and extrapolation uncertainty

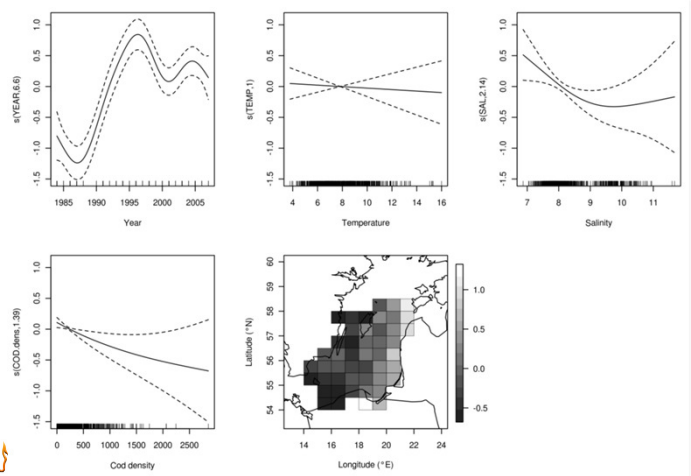
- a message from analogy predictions supporting management advice

Ullrika.Sahlin@lnu.se PhD

"Metaphorically, judgment is a kind of intellectual glue, cementing together the evidence and the methods"

Weed "**Weight of Evidence: A Review of Concept and Methods" Risk Analysis** *Vol. 25, No. 6, 2005*

Extrapolation uncertainty



CADASTER

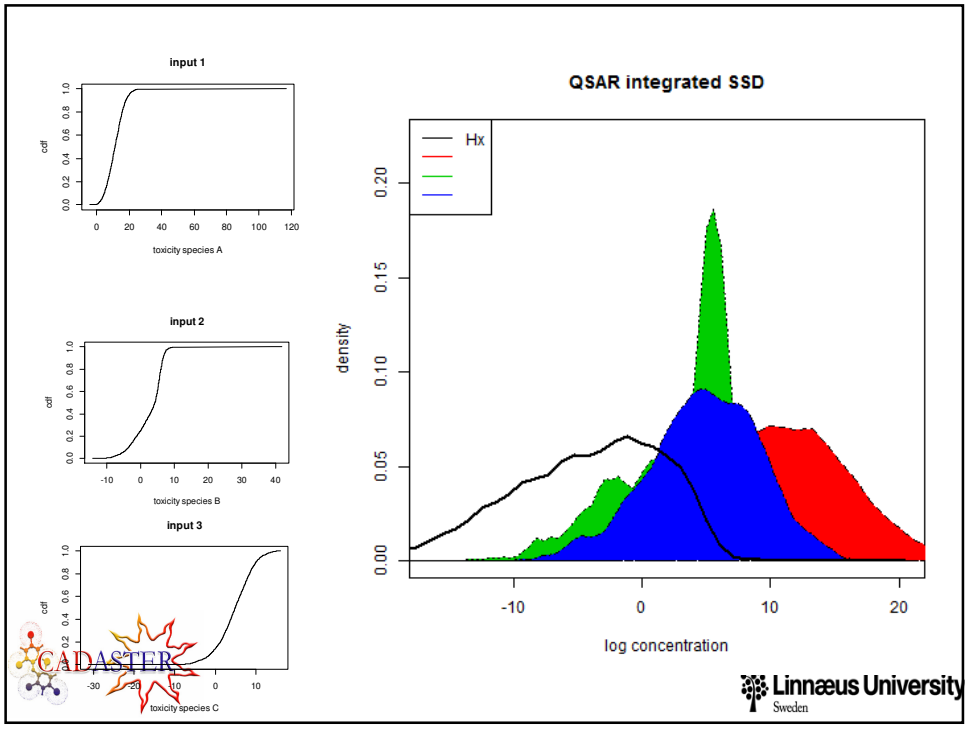
Linnæus University Sweden

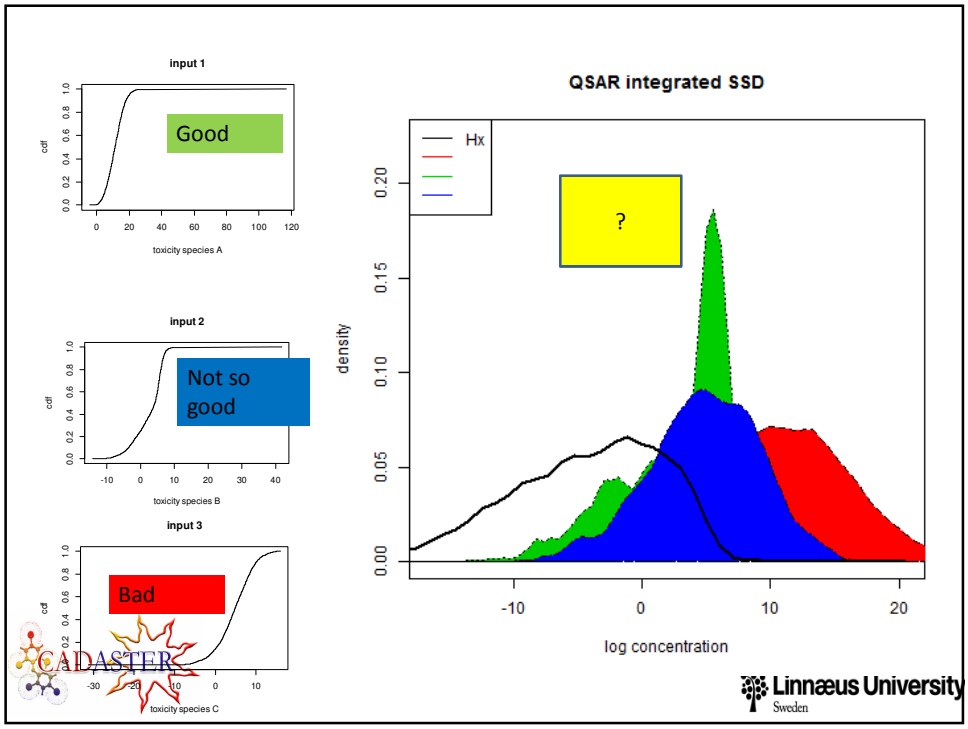
Uncertainty and sensitivity analyses

- How can we show when to get worried that an assessment is not good enough to support a decision?

CADASTER

Linnæus University Sweden


input 1
cdf
toxicity species A
input 2
cdf
toxicity species B
input 3
cdf
toxicity species C
QSAR integrated SSD
Hx
density
log concentration
Linnæus University
Sweden

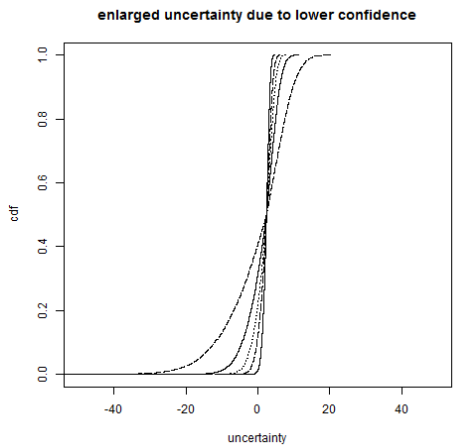

input 1
Good
cdf
toxicity species A
input 2
Not so good
cdf
toxicity species B
input 3
Bad
cdf
toxicity species C
QSAR integrated SSD
Hx
?
density
log concentration
Linnæus University
Sweden

Info-quality analysis

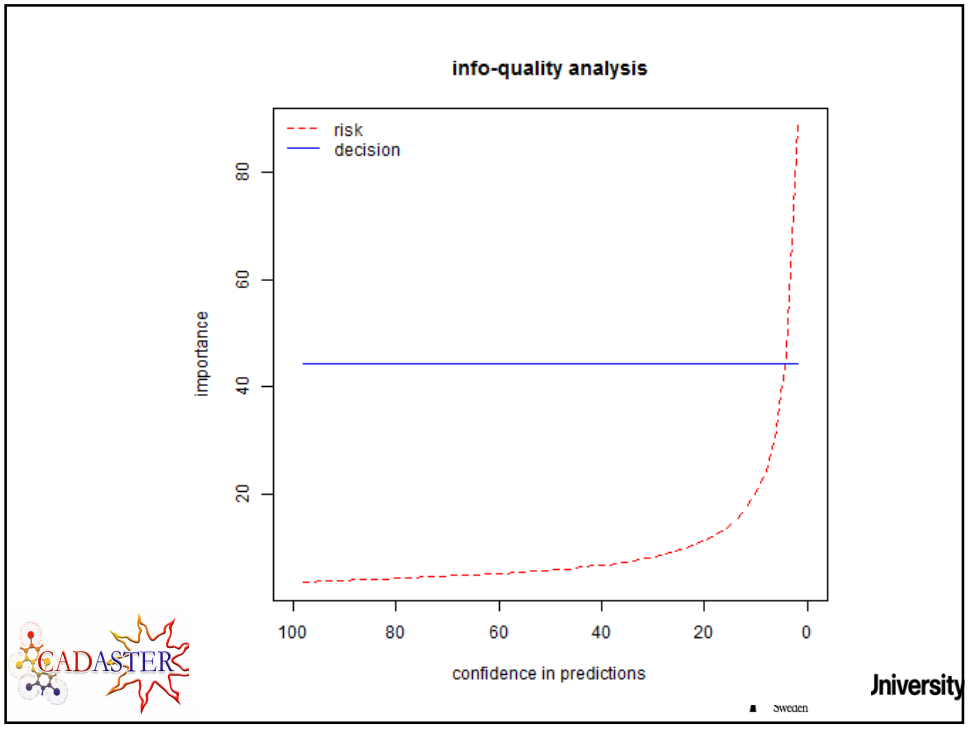
- A suggestion on how to propagate lower confidence in background knoweldge in an assessment
  - Quality is context dependent
  - Relate quality in background knowledge to confidence
  - and confidene to quality in decision support

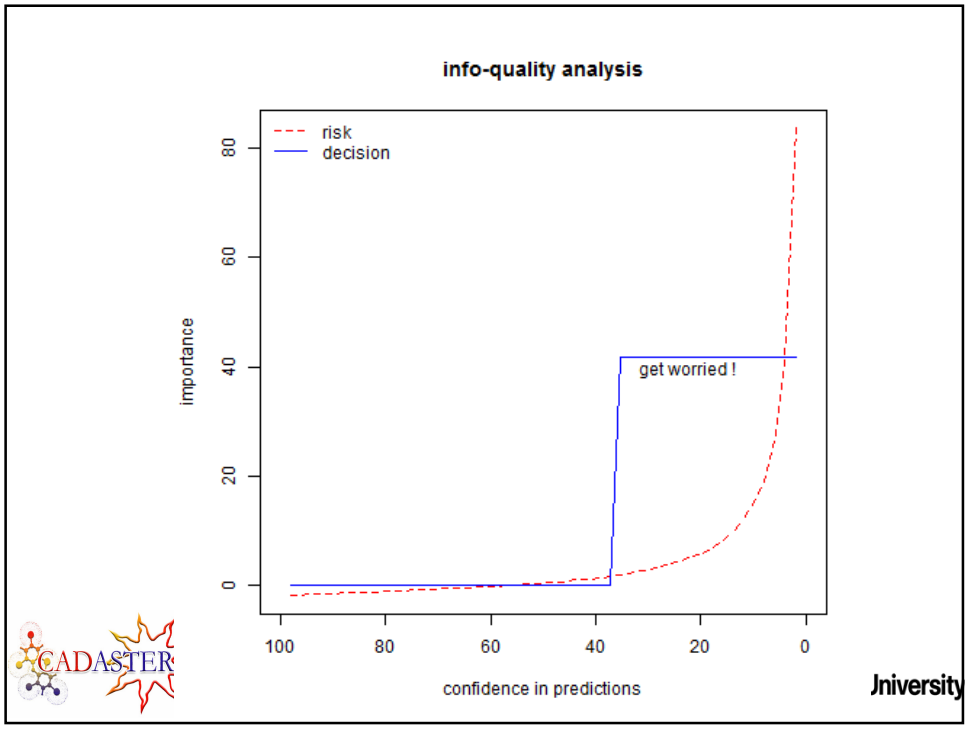
Linnæus University Sweden

Enlarge unc to reflect low confidence



Linnæus University Sweden


info-quality analysis
risk
decision
importance
20
40
60
80
100
80
60
40
20
0
confidence in predictions
CADASTER
Jniversity
Sweden


info-quality analysis
risk
decision
get worried !
importance
0
20
40
60
80
100
80
60
40
20
0
confidence in predictions
CADASTER
Jniversity

Analogy predictions

”Similar items ougth to have similar properties or behaviour”

Species traits

Habitat suitability

Climate matching

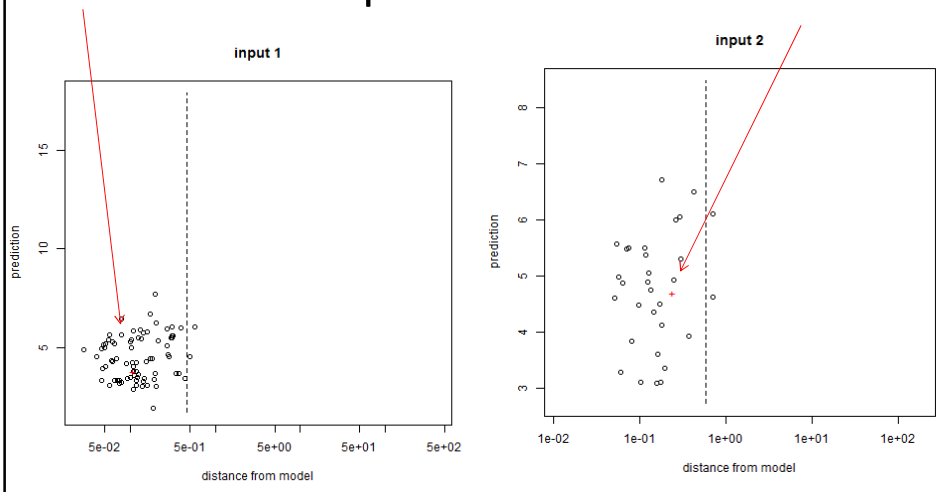
Historical events

Chemical structure

CADASTER

Linnæus University Sweden

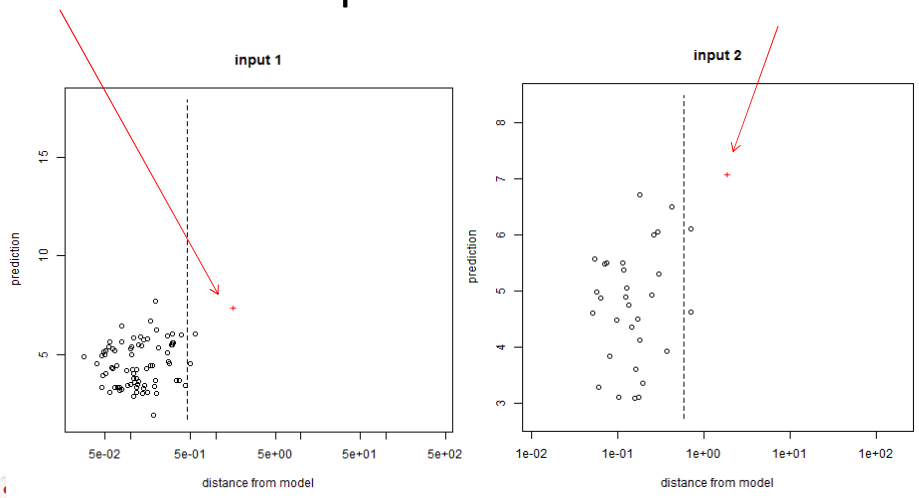
Extrapolation uncertainty in analogy predictions



CADASTER

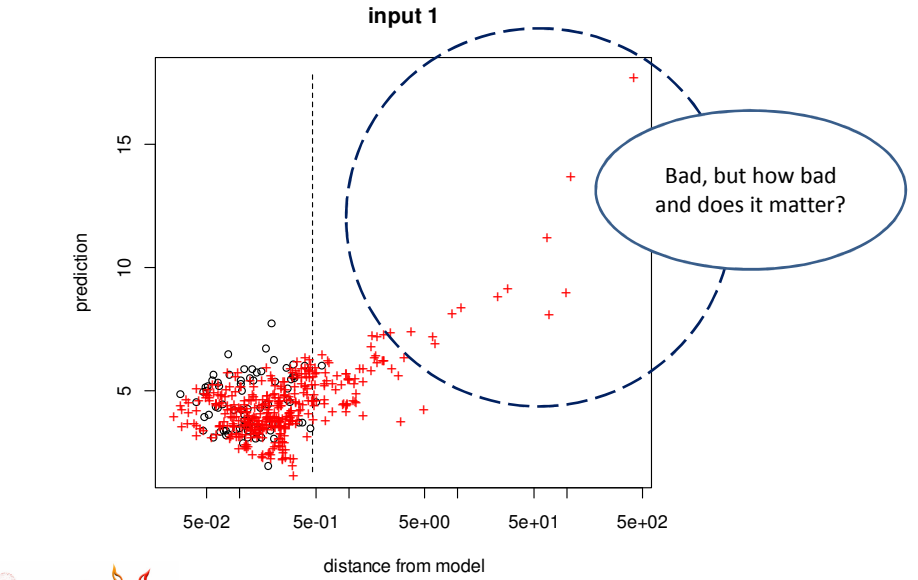
Linnæus University Sweden

Extrapolation uncertainty in analogy predictions



CADASTER

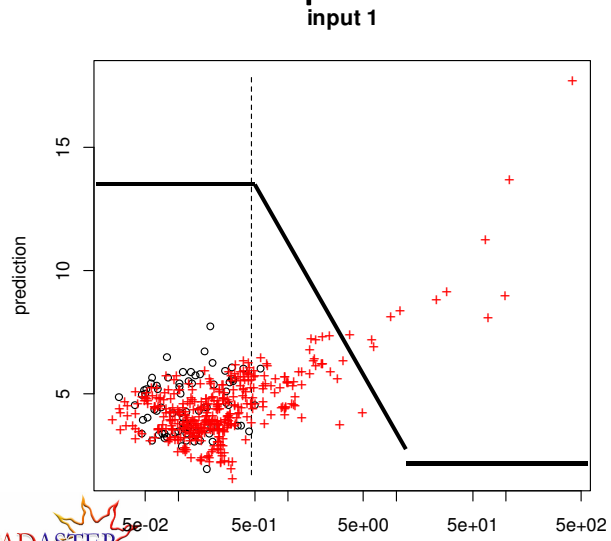
Linnæus University Sweden



CADASTER

Linnæus University Sweden

Relate extrapolation to confidence

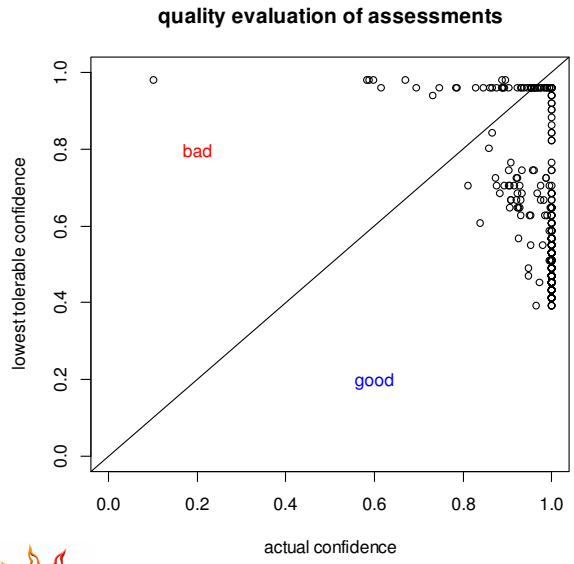


Highest confidence

Lowest confidence

CADASTER

Linnæus University Sweden



CADASTER

Linnæus University Sweden

Quick summary

- Scientific advice is evalutated in the step - Managerial review and judgment
- We can assure quality by sensitivity analysis towards different kinds of sources of uncertainty
- Extrapolation uncertainty cannot be statistically quantified, violates assumption of exchangeability, and exists in many assessments
- Info-quality analysis was here examplified with assessments based on analogy predicions

CADASTER

Linnæus University Sweden

Appendix 7

QSAR-integrated SSD modeling

Ullrika Sahlin (corresponding author) ullrika.sahlin@lnu.se

Linneaus University, Sweden

Introduction

The Species Sensitivity Distribution (SSD) approach (Posthuma II, Suter et al. 2002)(Posthuma II, Suter et al. 2002)(Posthuma II, Suter et al. 2002) to assess the Predicted No Effect Concentration (PNEC) in the environment treats sensitivities of observed species as random sample from the ecosystem. At least 10 species, spread over at minimum 8 taxonomic groups, are required for a proper SSD. The SSD approach is to fit a SSD to log EC50 values and derive the hazardous concentration considering uncertainty from sample size (Aldenberg and Jaworska 2000)(Aldenberg and Jaworska 2000)(Aldenberg and Jaworska 2000). Uncertainty in the hazardous concentration based on point predictions are derived as the non-central Student-t distribution. The PNEC value is then the median of the distribution for the hazardous concentration.

Experimental Species Data is SSDs are usually taken as fixed, which means that within species variability is usually not considered. Individual species data can vary in both in measurement uncertainty, or vary in quality as judged by experts, which may be taking into account to improve hazard or effect assessment (see e.g. O'Hagan, Craney et al. 2005)(see e.g. O'Hagan, Craney et al. 2005)(see e.g. O'Hagan, Craney et al. 2005). The added error from using a QSAR prediction compared to the experimentally based estimate ought to be considered when combining these kinds of information. Here we are concerned with what happens when we are to consider uncertainty in QSAR predictions, and in a mixture with experimentally established estimates of species sensitivities. We seek an approach to consider uncertainty in QSAR predictions that do not reduce variability in species sensitivities.

First, we conclude that there is a need of some sort of hierarchical modeling. The top level is the PNEC which is a statistical property of the SSD, such as the 5^{th} percentile (Aldenberg and Jaworska 2000)(Aldenberg and Jaworska 2000)(Aldenberg and Jaworska 2000). The second level is the species sensitivity which is a random variable with a defined probability distribution (e.g. lognormal). The third level is the individual species sensitivities, each associated to a certain amount of uncertainty. Here species sensitivity is an EC50 values, which means that within species variability is not modeled. Second, we conclude that the derivation of the hazardous concentration is not made through inference from data via a probability model since the predictive distributions associated to each QSAR predictions are initially given. A straightforward approach is to perform an outer loop using Monte Carlo simulation to sample from the predictive distributions and storing the hazardous concentration in each iteration.

Method

The influence of uncertainty in QSAR predictions on the PNEC was studied in a simulation experiment. Artificial data sets of species sensitivities were created and PNEC values derived by Monte Carlo simulation from distributions describing uncertainty in species sensitivities. PNEC was the hazardous concentration defined as the median value of the non-central Student-t distribution, assuming the underlying SSD to be normal.

$$SSD \sim N(\mu, \sigma) \qquad \text{Eq 1}$$

The sensitivity of an individual species to the chemical in concern T_i (in log units) is assumed to follow a normal distribution,

$$T_i \sim N(m_i, s_i) \qquad \text{Eq 2}$$

where i = 1, …, n is the number of species.

Artificial data sets were created based on an assumed SSD. Median values for each species are created by sampling from the SSD for a given SSD mean (μ) and SSD variation (σ^2). A species sensitivity described by an experimental value have here no uncertainty and therefore the variance is zero (i.e. s_i = 0 in Eq 2). The amount of QSAR predictions among the sample of n species is defined by k. Uncertainty in a species sensitivity represented by a QSAR prediction is assigned equal sized variances (i.e. $s_i = s$ in Eq 2).

An artificial data set is characterized by the parameters μ, σ, k and s, and were created for n = 10 species for the purpose of this experiment. The mean value of the SSD was kept constant (μ = 0), as it does not affect the influence of sources of variation. Latin Hypercube sampling was performed to generate 1000 parameter values within specified ranges (Table 1). An artificial SSD data set is illustrated in Figure 1 where five species are represented by QSAR predictions having the same variance. The theoretical SSD and hazardous concentration are shown in black.

Table 1. Specifications for the artificial SSD data sets consisting of 10 species.

Parameter	Description	Range
σ	SSD standard deviation	(0.01, 4)
k	Number of QSAR predictions	0,1,…,10
s	Standard deviation in QSAR prediction	(0.01,4)

For each sample the hazardous concentration was derived with and without considering QSAR uncertainty. The hazardous concentration was defined as the median of the 5th percentile in a SSD, where the median was taken from the non-central Student-t as described in (Aldenberg and Jaworska 2000)(Aldenberg and Jaworska 2000)(Aldenberg and Jaworska 2000). The hazardous concentrations with and without considering QSAR uncertainty corresponds to the blue and red dotted lines in Figure 1. In the case where QSAR uncertainty is considered, this median value is uncertain in itself as the QSAR uncertainty adds variation on another level (as a second order uncertainty).

The performance of the method to consider QSAR uncertainty was calculated as the reduction in the hazardous concentration on a log scale. The comparison was made between the mean value of the second order uncertainty to compare with the hazardous concentration when uncertainty in QSAR predictions is not considered.

All calculations were done in R (R Development Core Team 2008) and the code is available in Appendix 1.

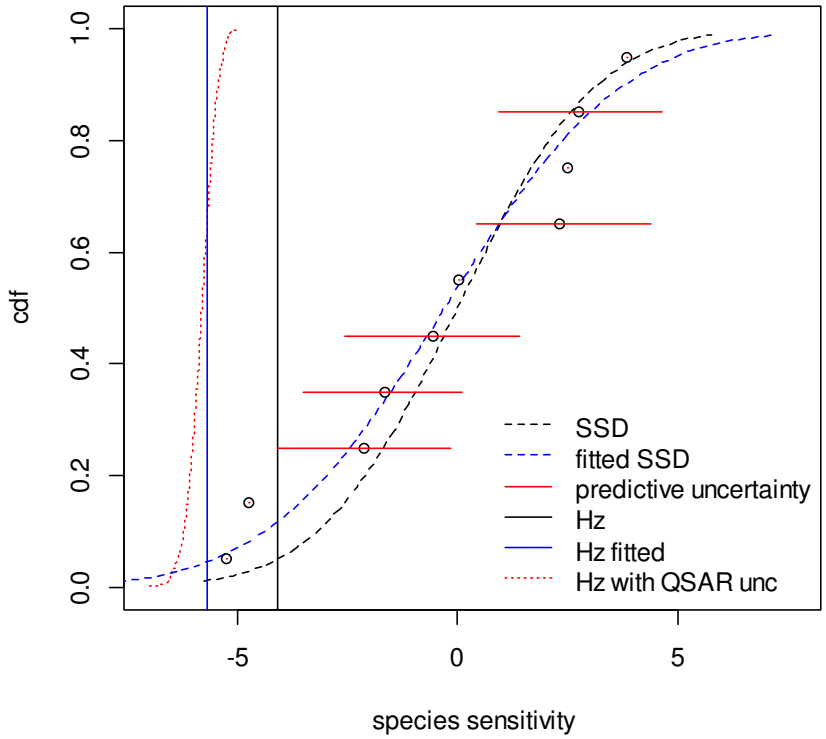


Figure 1. An illustration of the QSAR integrated SSD approach.

Results and Discussion

The simulation experiment showed that considering QSAR uncertainty results in a lower hazardous concentration, seen as positive performance measures (Figure 2). The influence of QSAR uncertainty is reduced with an increasing SSD variability (Figure 2 a), and increases with increasing number of species with QSAR predictions (Figure 2 b) and magnitude of uncertainty in QSAR predictions (Figure 2 c). Besides these rather intuitive results, we can confirm that there was no reduction in SSD variability from this way to consider uncertainty in QSAR predictions, i.e. by MC-sampling from predictive distributions as an outer loop of the standard SSD approach.

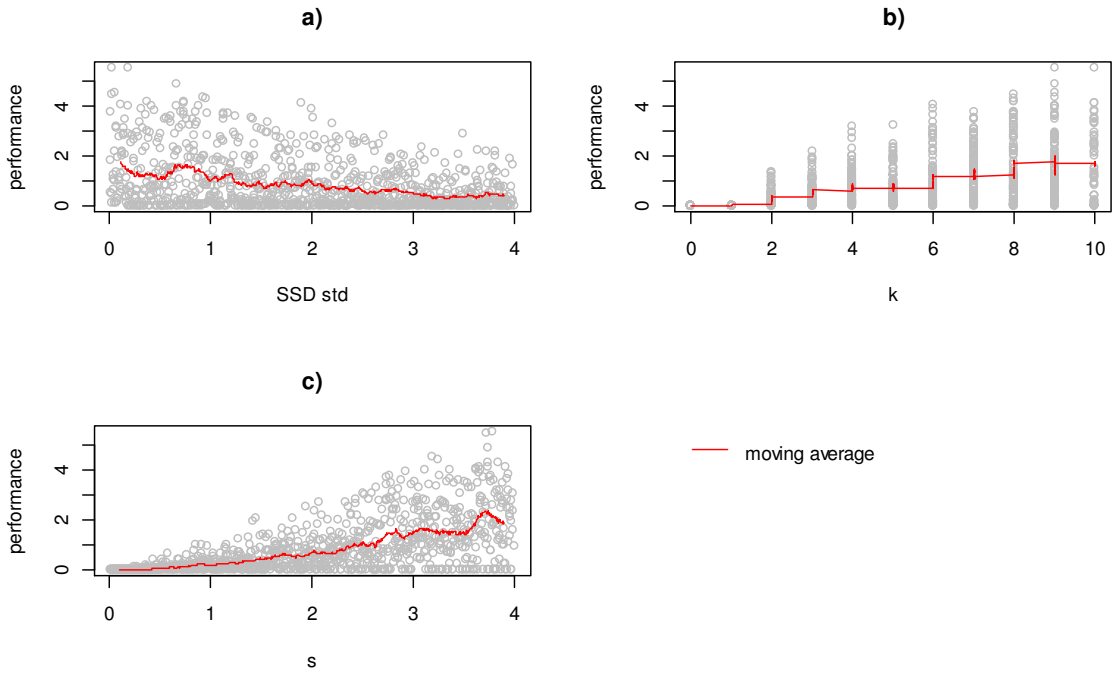


Figure 2. Results from the simulation experiment studying the difference between the hazardous concentration (median of the 5th percentile in a SSD) with and without considering uncertainty in QSAR predictions.

When the variance in predictive distributions for QSAR predictions are equal or smaller than the variance of the SSD, the change in hazardous concentration when considering QSAR uncertainty is within two order of magnitude (Figure 3). This difference is larger the more species with QSAR predictions in the SSD data set. A simple rule of thumb to consider QSAR uncertainty could be to divide the hazardous concentration derived from a SSD derived without considering QSAR uncertainty by a factor of 10 (if unlogged, or subtract 1 if on a log scale). A problem is that we do not know the variability in the SSD and that can only be determined on the available SSD data set, out of which the QSAR predictions are part. The procedure could be to fit an SSD without considering uncertainty in QSAR predictions and compare the variance of the fitted SSD to the average variance of the QSAR predictions, to judge how large uncertainty factor that is needed. Another option is to assess the hazardous concentration considering QSAR uncertainty directly. In Figure 2 we see that the influence from QSAR uncertainty is negligible when the variance in the QSAR predictive distribution is less than 10 % of the variance in the SSD.

Conclusions

Considering QSAR uncertainty in SSD approach in the assessment of PNEC is possible by keeping uncertainty in QSAR predictions and SSD variability at different levels comparable to second order Monte Carlo simulations.

Considering QSAR uncertainty generates more conservative hazardous concentrations.

When the average variance in QSAR predictions are less than the variance of the SSD, the added conservatism in the assessment of the hazardous concentration correspond to an assessment factor of less than 10.

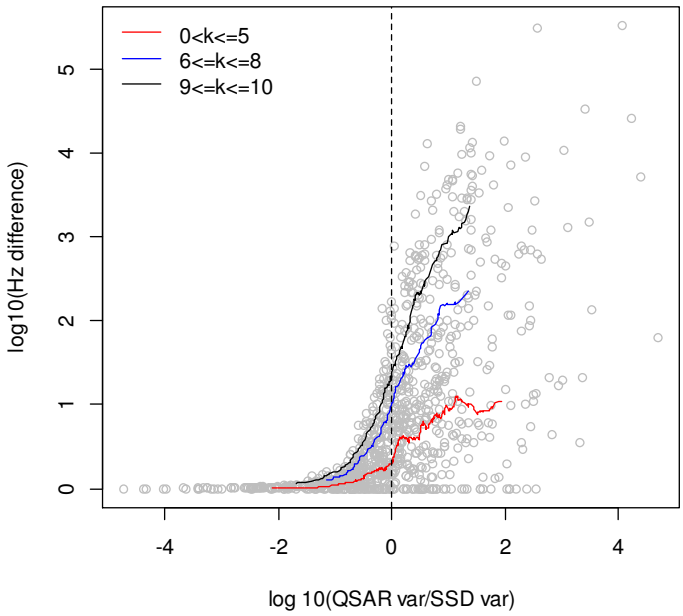


Figure 3. The relation between change in hazardous concentration (Hz) when considering QSAR uncertainty and how the magnitude of QSAR uncertainty (QSAR var) relates to the variability in the SSD (SSD var), shown for different number of species represented by QSAR predictions among 10 species in a SSD data set (k).

References

Aldenberg, T. and J. S. Jaworska (2000). "Uncertainty of the hazardous concentration and fraction affected for normal species sensitivity distributions." Ecotoxicology and Environmental Safety 46(1): 1-18.

O'Hagan, A., M. Craney, et al. (2005). Estimating species sensitivity distributions with the aid of expert judgements. Research Report No. 556/05; Department of Probability and Statistics, University of Sheffield: Sheffield, U.K., 2005. Freely downloadable from http://www.shef.ac.uk/st1ao/pub.html.

Posthuma II, L., G. W. Suter, et al. (2002). Species Sensitivity Distributions in Ecotoxicology. Boca Raton, Lewis.

R Development Core Team (2008). " R: A language and Environment for Statistical Computing." R Foundation for Statistical Computing.

Appendix 1

```
#install.packages('lhs')
library(lhs)

setwd('C:/Users/ulsaaa/Dropbox/wp4_summer/tom/SSDtoDeliverable')

## set up simulation experiment and generate artificial SSD data
n <- 10
mu <- 0
sigma_val <- c(0.01,4)
k_val <- c(0,10)
s_val <- c(0.01,4)

minmax <- t(cbind(sigma_val,k_val,s_val))

nLH <- 1000
LH <- randomLHS(n=nLH, k=3)
for(i in 1:3){
LH[,i]<-(minmax[i,2]-minmax[i,1])*LH[,i]+minmax[i,1]
}
LH[,2] <- round(LH[,2])



## function for QSAR integrated SSD
QSARSSD <- function(sample = array(rnorm(100*3),c(100,3)),x = 5,dist =
'non.central.t',non.central.t.arg = 50){
## x% is the percentile
## sample is a matrix of samples from diferent species (columns)
## dist = c('normal','t','non.central.t')
## the use on non.central.t requires an additional argument which is the percentle of the resulting
distribution given in percent

if(dist == 'normal'){
Hx <- qnorm(x/100,apply(sample,1,'mean'),sqrt(apply(sample,1,'var')))
}else if(dist == 't'){
Hx <- qt(x/100,dim(sample)[2])*sqrt(apply(sample,1,'var'))+apply(sample,1,'mean')
}else if(dist =='non.central.t'){
##it provides the non.central.t.arg percentile of the Hx seen over unc from small sample size in the
model
##calculate directly and follows tabulated values in Aldenberg and Jaworska 2000
kp <- -qnorm(x/100)
n <- dim(sample)[2]
ks <- qt(non.central.t.arg/100,n-1,ncp = kp*sqrt(n))/sqrt(n)
Hx <- apply(sample,1,'mean')-ks*sqrt(apply(sample,1,'var'))
}
#option of other distributions here
#H.out <- cbind(Hx,sample)
#return(H.out)
return(Hx)
}
```

```
N <- 1000
Hx <- array(0,c(N,nLH,2))
sample <- array(0,c(N,n))
sample.nounc <- sample
for(ind in 1:nLH){
#artifical
for(i in 1:n){## n is the number of species
sample.nounc[,i] <- array(rnorm(1,0,LH[ind,1]),c(N,1))
if(i > LH[ind,2]){
sample[,i] <- sample.nounc[,i]
}else{
sample[,i] <- sample.nounc[,i]+rnorm(N)*LH[ind,3]
}#end if
}#end i
Hx[,ind,1] <- QSARSSD(sample)
Hx[,ind,2] <- QSARSSD(sample.nounc)
}#end ind

### Performance measure
per <- apply(Hx[,,2]-Hx[,,1],2,'mean')

### Figure 2
k <- 50
par(mfrow = c(2,2))
plot(LH[,1],per,xlab = 'SSD std',ylab = 'performance',col = 'gray',main = 'a)')
MA <- filter(per[sort.int(LH[,1],index = TRUE)$ix],rep(1, k),method = "convolution",sides = 2)/k
lines(sort(LH[,1]),MA,col = 'red')

plot(LH[,2],per,xlab = 'k',ylab = 'performance',col = 'gray',main = 'b)')
MA <- filter(per[sort.int(LH[,2],index = TRUE)$ix],rep(1, k),method = "convolution",sides = 2)/k
lines(sort(LH[,2]),MA,col = 'red')

plot(LH[,3],per,xlab = 's',ylab = 'performance',col = 'gray',main = 'c)')
MA <- filter(per[sort.int(LH[,3],index = TRUE)$ix],rep(1, k),method = "convolution",sides = 2)/k
lines(sort(LH[,3]),MA,col = 'red')

plot(c(0,1),c(0,1),col = 'white',axes = FALSE,xlab = '',ylab = '')
legend('topleft','moving average',col = 'red',lty = 1,bty = 'n')

### Figure 3
LHtemp <- log10(LH[,3]^2/LH[,1]^2)
sel <- LH[,2]>0
temp <- LHtemp[sel]
pertemp <- per[sel]
plot(temp,pertemp,xlab = 'log 10(QSAR var/SSD var)',ylab = 'log10(Hz difference)',col = 'gray')

sel <- (LH[,2]>0 & LH[,2]<=5)
temp <- LHtemp[sel]
```

```
pertemp <- per[sel]
MA <- filter(pertemp[sort.int(temp,index = TRUE)$ix],rep(1, k),method = "convolution",sides = 2)/k
lines(sort(temp),MA,col = 'red')

sel <-  (LH[,2]>5 & LH[,2]<8)
temp <- LHtemp[sel]
pertemp <- per[sel]
MA <- filter(pertemp[sort.int(temp,index = TRUE)$ix],rep(1, k),method = "convolution",sides = 2)/k
lines(sort(temp),MA,col = 'blue')

sel <-  LH[,2]>=8
temp <- LHtemp[sel]
pertemp <- per[sel]
MA <- filter(pertemp[sort.int(temp,index = TRUE)$ix],rep(1, k),method = "convolution",sides = 2)/k
lines(sort(temp),MA,col = 'black')
legend('topleft',c('0<k<=5','6<=k<=8','9<=k<=10'),col = c('red','blue','black'),lty = c(1,1,1),bty = 'n')
abline(v = 0,lty = 2)

### test the effect of each factor
fit <- lm(per ~ poly(LH, 2, raw=TRUE))
summary(fit)
fit <- lm(per ~ LHtemp+LH[,2])
summary(fit)




#### Figure 1

ind <-12
n <- 10
sample <- array(0,c(N,n))
sample.nounc <- sample
for(i in 1:n){## n is the number of species
sample.nounc[,i] <- array(rnorm(1,0,LH[ind,1]),c(N,1))
if(i > LH[ind,2]){
sample[,i] <- sample.nounc[,i]
}else{
sample[,i] <- sample.nounc[,i]+rnorm(N)*LH[ind,3]
}#end if
}#end i
Hx <- QSARSSD(sample,non.central.t.arg = 50)
Hx.nounc <- QSARSSD(t(array(apply(sample,2,'mean'),c(n,2))),non.central.t.arg = 50)

sind <- sort.int(apply(sample,2,'mean'),index.return = TRUE)$ix
plot(range(Hx,sample,qnorm(0.01,0,LH[ind,1]),qnorm(0.999,0,LH[ind,1])),c(0,1),col = 'white',xlab =
'species sensitivity',ylab = 'cdf')

lines(sort(Hx),(1:length(Hx))/length(Hx),col = 'red',lty = 3)
lines(qnorm(seq(0.01,0.99,0.01),0,LH[ind,1]),seq(0.01,0.99,0.01),lty = 2,col = 'black')
```

```
lines(qnorm(seq(0.01,0.99,0.01),mean(apply(sample,2,'mean')),sqrt(var(apply(sample,2,'mean')) )),
seq(0.01,0.99,0.01),col = 'blue',lty = 2)

abline(v = Hx.nounc[1],col = 'blue',lty = 1)
abline(v = qnorm(0.05,0,LH[ind,1]),col = 'black',lty = 1)
points(sort(apply(sample,2,'mean')),(1:n)/n-0.5/n)
for(i in 1:n){
lines(quantile(sample[,sind[i]],probs = c(0.025,0.975)),rep(i-0.5,2)/n,col = 'red',lty = 1)
}

legend('bottomright',c('SSD','fitted SSD','predictive uncertainty','Hz','Hz fitted','Hz with QSAR unc'),
col = c('black','blue','red','black','blue','red'),lty = c(2,2,1,1,1,3),bty = 'n')
```