Distr.
GENERAL

CES/AC.68/2002/10
18 February 2002

RUSSIAN
Original: ENGLISH

СТАТИСТИЧЕСКАЯ КОМИССИЯ и ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ

КОНФЕРЕНЦИЯ ЕВРОПЕЙСКИХ СТАТИСТИКОВ

ОРГАНИЗАЦИЯ ЭКОНОМИЧЕСКОГО СОТРУДНИЧЕСТВА И РАЗВИТИЯ (ОЭСР)

КОМИССИЯ ЕВРОПЕЙСКИХ СООБЩЕСТВ (ЕВРОСТАТ)

Совместное совещание ЕЭК/Евростата/ОЭСР
по национальным счетам
(Женева, 24-26 апреля 2002 года)

ОБНАРУЖЕНИЕ ВЫБРОСОВ МНОГОМЕРНЫХ ЭКСТРАПОЛИРОВАННЫХ АДМИНИСТРАТИВНЫХ ДАННЫХ В НАЦИОНАЛЬНЫХ СЧЕТАХ БЕЛЬГИИ

Специальный документ, представленный Национальным банком Бельгии*

Введение

1. Цель настоящего документа заключается в представлении основных результатов системы обнаружения выбросов, используемой в использовании в национальных счетах Бельгии применительно к основным переменным счета производства сектора S11. Главное внимание уделяется результатам, а не теоретическим выводам.

2. Оценка переменных выпуска (P1), промежуточного потребления (P2) и добавленной стоимости (B1g) сектора S11 в значительной степени (20% совокупной добавленной стоимости) опирается на экстраполяцию частичных административных данных. Поскольку экстраполирование данных всегда чревато риском ошибочной экстраполяции, существует потребность в надежной и предпочтительно автоматизированной системе обнаружения выбросов. Поскольку рассматриваемые переменные являются взаимосвязанными, предпочтение отдается многомерному анализу.

---

* Автор: г-н Мишель Судан, Отдел статистики Национального банка Бельгии.

3. Структура настоящего документа следует простой логике. Сначала в нем приводится справочная информация о бельгийской системе оценки, после чего кратко описываются рассматриваемые данные с использованием некоторых базисных статистических данных и анализа главных компонентов. Затем поясняется использование расстояния Махаланобиса для обнаружения выбросов. Последняя часть посвящена описанию внедрения этой концепции в национальных счетах Бельгии.

Справочная информация

4. Оценки переменных P1 и P2 сектора S11 рассчитываются с использованием квазивсеобъемлющих административных источников и обследований. Наиболее важными из этих источников являются годовые отчеты компаний, декларации НДС, декларации взносов в фонды социального страхования и результаты структурных обследований предприятий.

5. Все экономические единицы классифицированы по институциональному сектору, отрасли, региону и категории. Исходные оценки P1, P2 и B1g сектора S11, следовательно, рассчитываются на весьма подробном уровне: по отрасли (249 отраслей), району (уровень III КТЕС, 44 района) и "категории" (10 категорий).

6. Район включен в качестве классификационной переменной с целью обеспечения максимальной согласованности между национальными и региональными счетами. Однако использование данной классификационной переменной с 2000 года привело к 19-кратному увеличению числа агрегатов: с примерно 1 300 до более 25 000. Вследствие этого проведение ручной проверки стало невозможным.

7. Категория описывает размер компании (крупные предприятия, малые и средние предприятия) и данные, которые имеются в административных источниках.
По некоторым категориям в наличии имеется информация только о добавленной стоимости или заработной плате. В случае других категорий в наличии имеется информация по переменным P1, P2 и B1g.

8. Если в наличии имеется информация только по одной из трех основных переменных счета производства, тогда для расчета оценок по всем трем переменным используется экстраполяция, опирающаяся на данные о схожей группе предприятий. Под схожей группой понимается, что входящие в ее состав предприятия классифицированы по тому же сектору, отрасли и региону. Они отличаются только по категории.

Исследование данных

9. Анализ выбросов должен учитывать два различных типа выбросов. Во-первых, P1 и P2 должны характеризоваться примерной линейной зависимостью (на уровне агрегатов), и эта взаимосвязь должна характеризоваться определенной стабильностью во времени. Любые изменения должны носить постепенный характер. Во-вторых, динамика агрегатов должна характеризоваться разумными пределами. Агрегаты, которые демонстрируют исключительно высокие или низкие темпы роста, вызывают подозрение.

10. Для облегчения возможности такого анализа в его охват должны быть включены обе переменные - P1 и P2. Для включения измерения "время" анализ должен охватывать переменные за более чем один год. Для осуществления многомерного анализа переменные должны быть сгруппированы в векторы.

11. Таким образом, изучаемая статистическая совокупность представляет собой набор полных, строго положительных векторов, которые содержат следующие переменные: $X = (P1_t\ P2_t\ P1_{t-1}\ P2_{t-1})'$

12. Для периода 1998-1999 годов X становится $(P1_{1999}\ P2_{1999}\ P1_{1998}\ P2_{1998})'$. B1g не включается в качестве переменной, поскольку может быть выведена непосредственно из P1 и P2. Совокупность содержит 23 432 вектора из возможных 28 390 векторов (83%). Разница между этими двумя цифрами объясняется тем, что индивидуальные характеристики являются нестабильными во времени. Так, некоторые комбинации (сектор, отрасль, регион, категория) могут существовать в один временной период, но отсутствовать в другой вследствие изменения классификации, слияний и приобретений или изменений в наличии данных. Кроме того, в 1999 году были созданы две новые категории, информация по которым отсутствовала в 1998 году. Это означает, что значительное число агрегатов по-прежнему требует ручной проверки, главным образом по соотношениям.

13. Вследствие громадного размаха значений этих переменных они подвергаются логарифмическому преобразованию. Наименьшим агрегатом совокупности является 34 евро, а наибольшим – 16,5 млрд. евро. Все переменные имеют логарифмическое представление (бельгийские франки). Как можно увидеть по диаграммам 1a – 1d, распределение преобразованных (и нормализованных) переменных представляется квазинормальным, поскольку в нем присутствует лишь легкая правая асимметрия. Однако критерий согласия Колмогорова–Смирнова опровергает гипотезу нормальности всех четырех переменных с однопроцентным пределом погрешности.

14. Таблицы 1 и 2 свидетельствуют о том, что переменные весьма схожи и обладают высокой корреляцией. Анализ главных компонентов с использованием матрицы корреляции показывает, что множество точек в четырехмерном пространстве может быть легко преобразовано в двух- и даже в одномерное пространство (см. таблицу 3).

15. Ассоциированные собственные векторы в таблице 4 свидетельствуют о том, что структура множества точек проста. Первым главным компонентом является просто переменная размера. Второй главный компонент - контраст между двумя рассматриваемыми годами. Другими компонентами являются дополнительные контрасты. Однако величина обусловливаемой ими вариации является незначительной.

16. Очевидно, что логарифмически преобразованные данные образуют множество точек в четырехмерном пространстве, которое принимает форму сильно вытянутого эллипсоида лишь с весьма небольшими отклонениями от его основной оси.

Метод обнаружения выбросов

17. Для проверки на выбросы по каждому наблюдению рассчитывается расстояние Махаланобиса. Концептуально этот процесс можно разбить на два этапа. Сначала рассматриваемые переменные трансформируются с использованием преобразования Махаланобиса, которое стандартизирует (среднее = 0, дисперсия = 1) и ортогонализирует несходные данные (т.е. их взаимные корреляции становятся равными нулю).

$$y = C^{-1/2}.(x - x_{avg}),$$

где

x = исходный вектор (логарифмически преобразованных) данных

y = вектор преобразованных данных

C = расчетная матрица ковариации

x_{avg} = расчетный вектор средних

18. Затем рассчитывается расстояние Махаланобиса между преобразованными данными и исходным 0 с использованием следующей формулы:

$$D^2 = y'.y = \sum y_i^2 = (x - x_{avg})'.C^{-1}.(x - x_{avg}),$$

где D ($= +\sqrt{D^2}$) равно расстоянию Махаланобиса.

19. Геометрически наблюдения с одинаковым расстоянием Махаланобиса формируют эллипсоид. Расстояние Махаланобиса является малым в случае наблюдений, находящихся на или вблизи от основной оси эллипсоида. Наблюдения более удаленных

от основной оси характеризуются намного большим расстоянием Махаланобиса. Диаграмма 2 иллюстрирует этот вывод применительно к двум переменным. Эллипсоид формируется всеми наблюдениями с расстоянием Махаланобиса равным 1. Расстояние Махаланобиса для (фиктивного) выброса (с координатами 7,5 и 8,5), означающего, что P2 в десять раз превышает P1, равно 7,12.

Реализация

20. Конечной целью системы обнаружения выбросов является выявление подозрительных наблюдений для того, чтобы они могли быть подвергнуты ручной проверке. Для этого общая совокупность наблюдений делится на две части в соответствии со средним размером наблюдения. Наблюдения, средняя величина четырех переменных которых превышает 1 млрд. бельгийских франков (примерно 25 млн. евро), называют крупными наблюдениями. Другие наблюдения называют малыми наблюдениями.

21. Представим, что имеется 2 558 крупных и 20 874 малых наблюдений. К наблюдениям, требующим ручной проверки, относятся наблюдения, для которых расстояние Махаланобиса находится в 5-процентном верхнем квантиле каждой группы. В случае крупных наблюдений это соответствует наблюдениям, для которых расстояние Махаланобиса составляет более 3. В случае малых наблюдений точкой раздела является 3,8. Таблица 5 содержит таблицы плотности распределения D, D_{large} и D_{small}.

22. В таблице 6 приводится выдержка из типичной таблицы выходных данных. Помимо исходных данных и значений самого расстояния Махаланобиса, в нее также включены преобразованные данные. Поскольку квадрат расстояния Махаланобиса равен сумме квадратов преобразованных переменных, преобразованные данные позволяют определить, какая переменная (переменные) из четырех, включенных в вектор, ответственна за значительное отклонение. Отметим, что Y1 является преобразованной переменной P1_98, Y2 - преобразованной переменной P1_99 и т.д. Преобразованные переменные выражаются в виде среднеквадратических отклонений.

23. Например, значительное расстояние Махаланобиса первого наблюдения обусловлено главным образом переменными P1_98 и P2_98 (Y1 и Y3, соответственно). Очевидно, что P1_98 является слишком низкой с учетом P2_98.

24. Выходные данные, приведенные в таблице 6, рассчитаны с использованием интерфейса компонентной объектной модели реализованного в макросах Excel. Это позволяет любому сотруднику Отдела статистики, имеющему доступ к Excel, генерировать выходные данные таблицы 6. В настоящее время ведется постепенное

внедрение дополнительных опций (матрицы ковариации и корреляции, графическое представление выходных данных и т.д.). Ранее для этого использовалась стандартная процедура SAS-IML, которая ограничивала возможности доступа.

Выводы

25. Хотя все наблюдения, которые были отобраны для ручной проверки в ходе рассмотренной выше работы, характеризуются отклонениями, в большинстве случаев после проведения проверки корректировка данных не проводится. Значительные колебания значений агрегатов на межгодовой основе обусловлены главным образом переходом компании в другую категорию, а не ошибочной экстраполяцией. С учетом опыта, накопленного в рамках использования расстояния Махаланобиса, отдел национальных счетов в настоящее время изучает различные способы обеспечения большей стабильности во времени классификационной переменной "категория".

26. Эксперименты с вышеописанным методом показали, что использование расстояния Махаланобиса является весьма эффективным способом обнаружения выбросов. Это весьма универсальный метод, поскольку его применение не ограничивается вышеописанным случаем. Другие потенциальные виды использования включают в себя сопоставление двух или более переменных, которые предположительно измеряют одно и то же явление (например, безработицу) или сопоставление двух вариантов переменных для выявления крупных корректировок.

27. Тем не менее существуют определенные ограничения. Наиболее важное из них - то, что переменные, включаемые в охват анализа, должны характеризоваться симметричным распределением и быть унимодальными. Корреляция между переменными должна быть весьма высокой (предпочтительно $|r|>90\%$). Зависимости между переменными должны быть линейными, хотя эта проблема может быть решена благодаря использованию линеаризующего преобразования. И наконец, рекомендуется использовать в анализе ограниченное число переменных (<10), чтобы не осложнять интерпретацию преобразованных *y*-переменных.

ТАБЛИЦЫ

Таблица 1. Исходные статистические данные

	P1_99	P2_99	P1_98	P2_98
среднее	7,83	7,64	7,80	7,63
среднеквадратическое отклонение	1,05	1,08	1,04	1,07
ассиметрия	-0,22	-0,19	-0,19	-0,19
эксцесс	0,15	0,06	0,08	0,06

Таблица 2. Матрица корреляции

	P1_99	P2_99	P1_98	P2_98
P1_99	100%			
P2_99	98,8%	100%		
P1_98	92,4%	91,6%	100%	
P2_98	91,7%	92,6%	98,8%	100%

Таблица 3. Анализ главных компонентов

Компоненты	Собственные значения	Соотношение	Совокупные значения
PRIN1	3,829	95,7%	95,7%
PRIN2	0,147	3,7%	99,4%
PRIN3	0,021	0,5%	99,9%
PRIN4	0,004	0,1%	100,0%

Таблица 4. Собственные векторы

	PRIN1	PRIN2	PRIN3	PRIN4
P1_99	0,500	-0,500	0,497	-0,503
P2_99	0,500	-0,500	-0,492	0,508
P1_98	0,500	0,505	0,504	0,492
P2_98	0,500	0,495	-0,508	-0,497

Таблица 5. Таблица плотности распределения

Bin	D		D_{large}		D_{small}	
	Freq	cum%	Freq	cum%	Freq	cum%
0	0	0%	0	0%	0	0%
1	8676	37%	569	22%	8107	39%
2	10109	80%	1613	85%	8496	80%
3	2757	92%	240	95%	2517	92%
4	961	96%	100	99%	861	96%
5	409	98%	8	99%	401	98%
6	199	99%	14	99%	185	99%
7	121	99%	4	100%	117	99%
8	64	99%	10	100%	54	99%
9	40	100%	0	100%	40	100%
10 и более	96	100%	0	100%	96	100%
□	23432		2558		2874	

Таблица 6. Выдержка из таблицы выходных данных (первые 10 строк)

P1_98	P1_99	P2_98	P2_99	lgP1_98	lgP1_99	lgP2_98	lgP2_99	Y1	Y2	Y3	Y4	Dist
5 000	115 492 000	10 001 000	121 220 000	3,70	8,06	7,00	8,08	-22,76	9,63	14,79	-3,93	29,07
17 312	1 527	5 330	818 299	4,24	3,18	3,73	5,91	5,76	-18,38	-10,36	16,03	27,12
20 998	12 389 988	87 354	49 274	4,32	7,09	4,94	4,69	-12,38	14,43	6,80	-13,46	24,27
295 365 000	185 706 000	30 417 000	241 000	8,47	8,27	7,48	5,38	-0,31	11,62	2,90	-14,50	18,81
43 112 339	162 874 745	20 907 989	589 571	7,63	8,21	7,32	5,77	-3,97	11,41	4,70	-12,86	18,26
27 863 551	5 350 146	117 642	2 799 390	7,45	6,73	5,07	6,45	10,76	-4,75	-12,58	4,24	17,73
300 828	809 454	1 392	430 287	5,48	5,91	3,14	5,63	8,32	-4,02	-14,28	4,96	17,72
905 405	13 127 648	188 538 380	39 897 241	5,96	7,12	8,28	7,60	-12,81	1,89	11,58	-1,86	17,47
54 000	7 000	2 299 000	2 463 000	4,73	3,85	6,36	6,39	-4,57	-13,22	3,26	9 50	17 22
17 569 618	1 249 355	15 723 584	15 563	7,24	6,10	7,20	4,19	-2,77	6,32	6,39	-12,75	15,85

Figure I.b
Histogram log(P1_99)

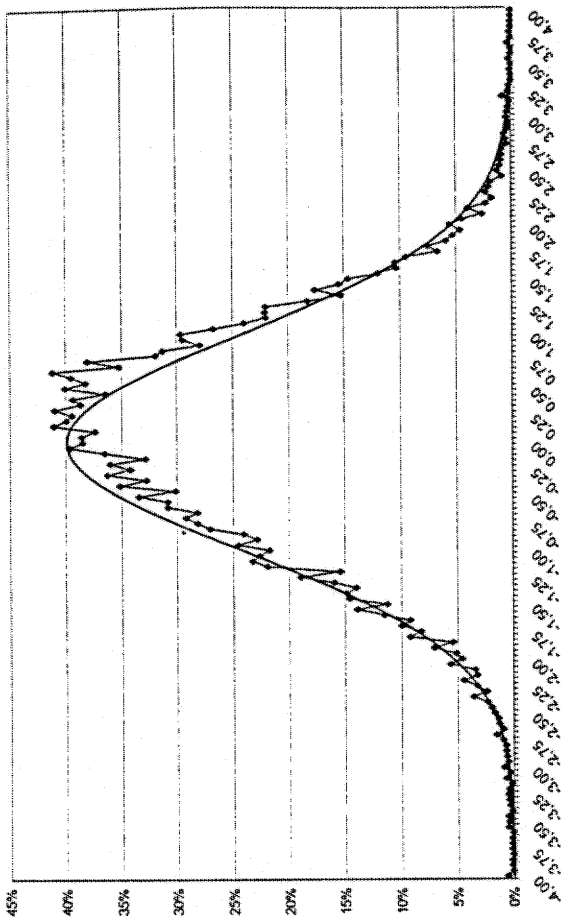


Figure I.c
Historgram log(P2_98)

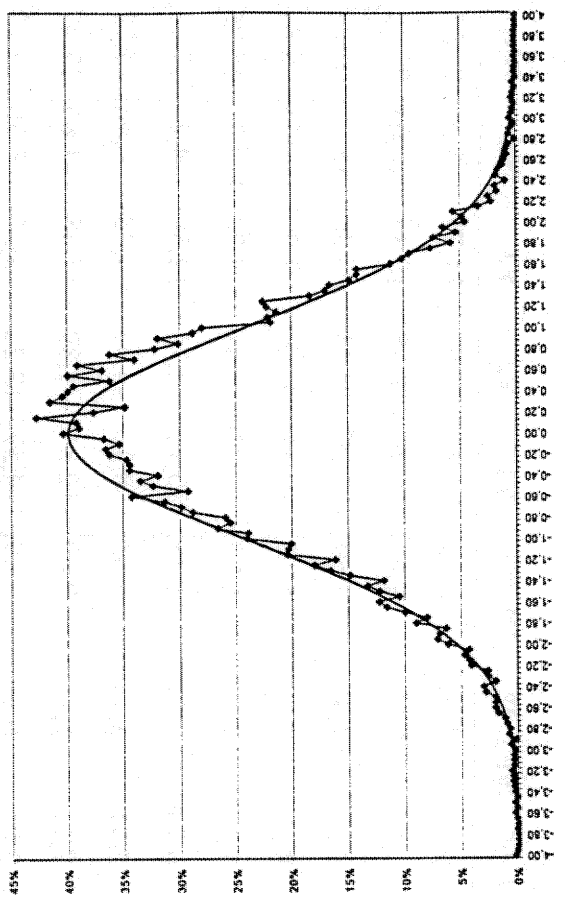


Figure I.a
Histogram log(P1_98)

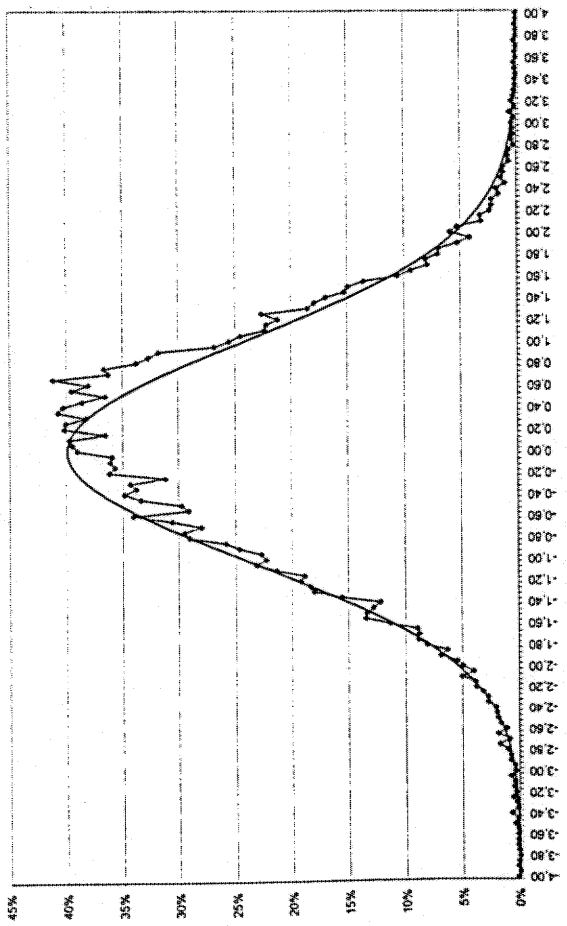
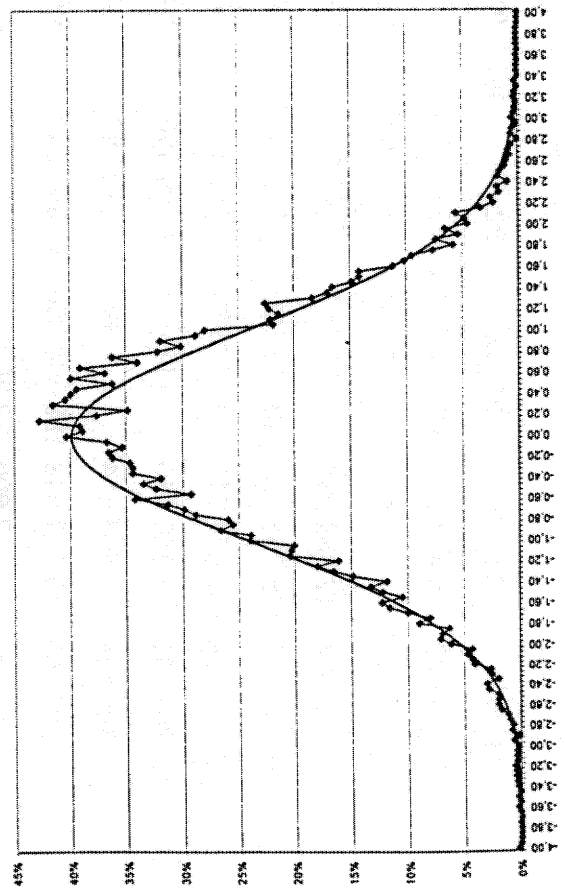


Figure I.c
Historgram log(P2_98)



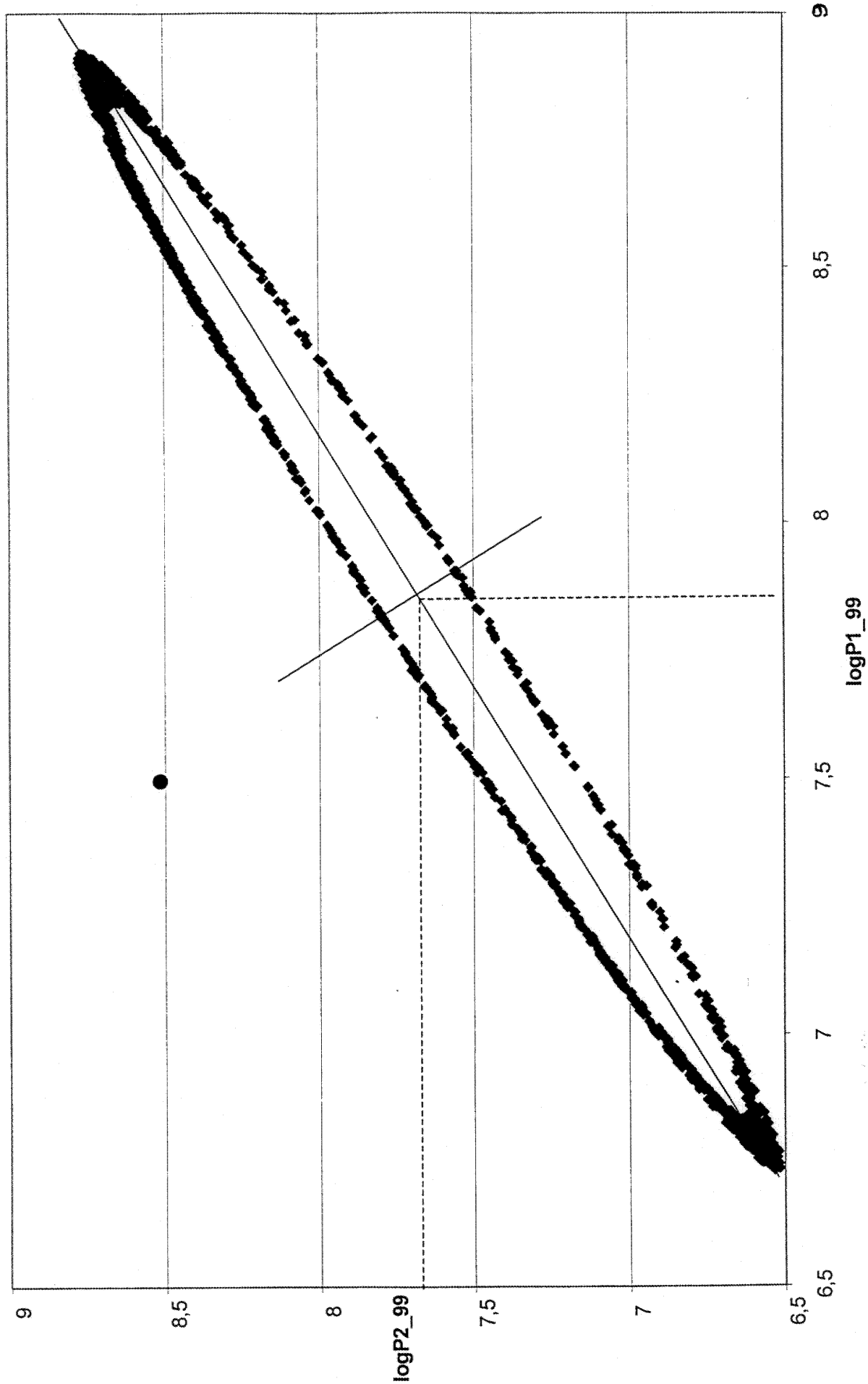


Figure 2: logP1_99 vs. logP2_99