Selektivitätsschätzung von Bereichsanfragen auf metrischen Attributen mit nichtparametrischen Verfahren

Dissertation

zur

Erlangung des Doktorgrades

der Naturwissenschaften

(Dr. rer. nat.)

dem

Fachbereich Mathematik und Informatik

der Philipps-Universität Marburg

vorgelegt von

Dieter Korus

aus Brilon

Marburg/Lahn 1999

Vom Fachbereich Mathematik und Informatik
der Philipps-Universität Marburg als Dissertation am 4. Februar 2000

angenommen.

Erstgutachter Prof. Dr. Bernhard Seeger

Zweitgutachter Prof. Dr. Manfred Sommer

Tag der mündlichen Prüfung am 11. Februar 2000

Vorwort

Datenbanksysteme sind heutzutage in jedem (größeren) Unternehmen vorzufinden. Die Datenmengen, die in Datenbanksystemen gespeichert werden, nehmen im gleichen Maße drastisch zu wie die Informationen, mit denen ein Unternehmen heutzutage überflutet wird. Insbesondere im Zusammenhang mit Data-Warehouse-Anwendungen werden bereits Terabyte-Datenbanksysteme diskutiert (z.B. teradata von NCR - [teradata], [Walter 98] - oder Terra-Server von Microsoft - [terra server]).

Mit zunehmender Größe der Datenbanken gewinnt auch die Schätzung der *Selektivität* einer Anfrage, d.h. die Schätzung der Größe der zu erwartenden Antwortmenge, vermehrt an Bedeutung. Eine wichtige Aufgabe für das Anfrageverarbeitungssystem eines Datenbank-Management-Systems ist die Bestimmung eines möglichst optimalen Ausführungsplans für komplexe Anfragen anhand von automatischen Selektivitätsschätzungen, um die Anfrage mit möglichst geringen Kosten bearbeiten zu können (*Anfrageoptimierung*). Gleichzeitig nimmt mit der Größe der Datenbanken aber auch deren Komplexität zu. Während es bei kleinen Datenbanken keine Rolle spielt, wie hoch die Kosten der Anfrage und wie gut die Qualität der Antwortmenge sind, muß der Benutzer großer Datenbanken zuvor vielmals eine Abschätzung der Kosten der Anfrage und der Qualität der Antwortmenge bekommen, um nicht unerwartet lange Antwortzeiten oder eine unübersichtliche Menge an Daten in Kauf nehmen zu müssen. Kosten und Qualität müssen gegeneinander abgewogen werden. Diese Kosten- und Qualitätsabschätzungen spielen eine Rolle beim sogenannten *Anfrageprofiling* in großen Datenbanksystemen, werden aber auch zunehmend im Rahmen von Data-Warehouse-Anwendungen und Data-Mining-Fragestellungen wichtig ([Fayyad et al. 96], [Glymour et al. 96], [Imielinski & Mannila 96], [Holsheimer & Kersten 94], [NCR 99]).

Eine weitere Tendenz im Zusammenhang mit Datenbanksystemen ist die Zunahme von Anwendungen, bei denen mehrdimensionale Indexstrukturen erforderlich sind. Beispiele hierfür sind Geo-Datenbanksysteme, aber auch bei der effizienten Speicherung und Abfrage von Bild- oder Textobjekten spielen diese Indexstrukturen eine wichtige Rolle ([Zezula et al. 96], [Hellerstein & Pfeffer 94]). Hierfür werden entsprechende Methoden zur Selektivitätsschätzung bei Relationen mit mehrdimensionalen Indexstrukturen gesucht.

In dieser Arbeit wird aufgezeigt, wie das Problem der Selektivitätsschätzung in großen Datenbanksystemen mit Methoden der nicht-parametrischen Statistik gelöst werden kann. Dabei wird sowohl auf die Schätzung der Selektivität bei Anfragen auf Relationen mit eindimensionalen als auch mehrdimensionalen Indexstrukturen eingegangen.

Im Gegensatz zu bisherigen Arbeiten zur Selektivitätsschätzung werden hier metrische Daten mit großem Wertebereich betrachtet. Metrische Daten sind Bestandteil von Geodatenbanksystemen, kommen aber auch häufig in anderen großen Datenbanken vor.

Ein wichtiges Kriterium für Selektivitätsschätzer ist, daß die Schätzung der Selektivität wesentlich geringere Kosten in Anspruch nimmt als die exakte Ermittlung der Selektivität bzw. die Beantwortung der Anfrage selbst. Außerdem kommen nur Verfahren in Betracht, die automa-

tisch und unüberwacht arbeiten. Somit scheiden z.B. Verfahren aus, die auf der Visualisierung von Daten beruhen.

Es existieren bisher eine Reihe von Methoden zur Selektivitätsschätzung in Datenbanksystemen, die jedoch eher auf einem ad hoc Vorgehen basieren. In der Literatur vorgestellte Arbeiten handeln zwar von Stichproben (Sampling), Histogrammen und "Gleichverteilungsannahmen", aber die Beziehung zwischen Selektivitätsschätzung und Schätzmethoden der mathematischen Statistik wurde nie klar herausgestellt. Lediglich in [Mannino et al. 88] wurde ein solcher Zusammenhang überhaupt konstatiert, der dort beschriebene Ansatz wurde aber nicht weiterverfolgt. Dagegen wird in dieser Arbeit der Begriff Selektivität mit Hilfe statistischer Begriffe definiert und die Selektivitätsschätzung klar in Beziehung zu Schätzmethoden der (nicht-parametrischen) Statistik gebracht. Damit lassen sich die bekannten bisher zur Selektivitätsschätzung verwandten Methoden eindeutig beschreiben sowie neue Verfahren entwickeln.

Das Problem der Selektivitätsschätzung wurde zuerst im Datenbanksystem System-R [Selinger et al. 79] adressiert. Dort wurde als Verteilung für alle im Datenbanksystem vorhandenen Datenbanken die Gleichverteilung der Daten angenommen. Es macht jedoch i.a. keinen Sinn parametrische Verfahren anzuwenden, da sie die Kenntnis der Verteilungsfamilie voraussetzen, die der wahren Verteilung einer Datenbank zugrundeliegt. Diese ist jedoch in Datenbanksystemen bei den meisten Datenbanken nicht bekannt. Im Anschluß an System-R entwickelte Verfahren benutzen daher Histogramme oder Stichproben zur Selektivitätsschätzung. Die dort vorgestellten Methoden weisen allerdings Einschränkungen an die zugrundeliegenden Daten auf. So werden z.B. in [Poosala et al. 96] lediglich nominale Daten oder solche mit extrem kleinem Wertebereich betrachtet. Die zunächst zur Selektivitätsschätzung in Datenbanksystemen mit eindimensionalen Indexstrukturen vorgestellten Methoden sind auch für Datenbanksysteme mit mehrdimensionalen Indexstrukturen erweitert worden (z.B. [Muralikrishna & DeWitt 88], [Poosala & Ioannidis 97]). Histogramme sind die in den meisten heutigen Datenbanksystemen eingesetzten Schätzverfahren (z.B. Oracle 8, Sybase). Allerdings wird der Datenbankadministrator in der Wahl sowohl der Stichprobengröße als auch der Anzahl Bins nicht vom System unterstützt.

In dieser Arbeit wird zum ersten Mal die Frage der Selektivitätsschätzung von Selektionsanfragen, insbesondere von Bereichsanfragen, systematisch in einen mathematischen Rahmen eingeordnet. Bisherige Verfahren zur Selektivitätsschätzung werden anhand von Kriterien der mathematischen Statistik klassifiziert und analysiert. Erweiterungen dieser nicht-parametrischen Verfahren zur Selektivitätsschätzung im Eindimensionalen (Häufigkeitspolygone, Average Shifted Histogramme) wie im Mehrdimensionalen (neue Partitionierverfahren für Histogramme) werden vorgestellt.

In den letzten Jahren erzielte ein völlig neues Verfahren zur nicht-parametrischen Dichteschätzung, die sogenannten Kernschätzer, die Aufmerksamkeit in der mathematischen Statistik ([Silverman 86], [Wand & Jones 95]). Sie zeichnen sich gegenüber Histogrammschätzern dadurch aus, daß sie stetig und unabhängig vom Startpunkt sind sowie ein besseres Konvergenzverhal-

ten besitzen. Kernschätzer erfordern lediglich einen gering höheren Berechnungsaufwand, haben dagegen aber den Vorteil, daß keine apriori Information über die Daten (z.B. in Form von Histogrammklassen) berechnet werden muß. Insbesondere weisen sie damit die notwendige Flexibilität bei dynamischen Datenbanken auf. In dieser Arbeit wird eine neue Methode zur Selektivitätsschätzung mit Hilfe von Kerndichteschätzern vorgestellt. Bei Kernselektivitätsschätzern hat sich zudem die Behandlung von Randeffekten als erforderlich und als erfolgreich erwiesen.

Viele in der Literatur verwendete Histogrammselektivitätsschätzer gehen davon aus, daß zur Bildung der Histogramme die gesamte Datenbank herangezogen wird. Dies macht bei der Größe der betrachteten Datenbanken keinen Sinn. Stattdessen wird in dieser Arbeit bei allen nicht-parametrischen Selektivitätsschätzverfahren vorausgesetzt, daß eine im Vergleich zur Kardinalität der gesamten Datenbank kleine Stichprobe vorliegt. Die Güte der Selektivitätsschätzung ist somit abhängig von der Stichprobengröße.

Ein in diesem Zusammenhang nicht zu vernachlässigendes Problem, das bei nicht-parametrischen Schätzern wie Histogramm- oder Kernschätzern auf Basis von Stichproben auftritt, ist die Wahl der Bandbreite. Sowohl eine zu hohe als auch eine zu niedrige Bandbreite führt zu größeren Fehlern. Dabei ist die optimale Bandbreite u.a. abhängig von der Stichprobengröße. Verschiedene in der Statistik bekannte Methoden der näherungsweisen Bestimmung einer optimalen Bandbreite werden in dieser Arbeit zur Selektivitätsschätzung angewendet und anhand von experimentellen Ergebnissen im ein- und zweidimensionalen Fall diskutiert.

Reale Daten zeigen oft ein anderes Verhalten als in der Theorie angenommen. Daher ist die Validierung der Verfahren mit realen Testdaten unerläßlich und ein wichtiger Bestandteil dieser Arbeit. Auch bei der Selektivitätsschätzung zeigt sich eine Diskrepanz der experimentellen Ergebnisse bei künstlichen und realen Daten. Die Ursache dafür ist, daß die in der Theorie vorausgesetzte Glattheit der Schätzfunktion bei den realen Daten i.a. nicht gegeben ist. Um das Problem zu lösen, wurde in dieser Arbeit der Hybridselektivitätsschätzer entwickelt. Er besteht aus einer Kombination von Histogramm- und Randkern-Selektivitätsschätzer. Hierbei wird der Datenraum anhand der Sprungstellen partitioniert und somit ein Histogramm erzeugt. Anschließend kann die lokale Selektivität einer Anfrage innerhalb der Histogrammbins durch Randkernselektivitätsschätzer geschätzt und zur Gesamtselektivität der Anfrage zusammengefaßt werden. Die experimentellen Ergebnisse bestätigen dieses Vorgehen.

Im Vorfeld dieser Arbeit sind erste Vorarbeiten in Form von Diplomarbeiten geleistet worden. Hierzu gehören die Vorstellung weiterer multivariater Selektivitätsschätzer auf Histogrammbasis ([Buskamp 97], [Schneider 97]) und die Verwendung von Kernschätzern zur Selektivitätsschätzung. ([Blohsfeld 98]) In der letzten Diplomarbeit wurden bereits erste Ansätze zur Bandbreitenbestimmung vorgestellt. Diese Diplomarbeiten sind wesentlich von mir betreut worden. Die dortigen Ergebnisse sind von mir in einen systematischen Zusammenhang gestellt, verbessert und fortgeführt worden. Teile dieser Arbeit im univariaten Fall sind vorab in [Blohsfeld et al. 99] veröffentlicht worden.

Die vorliegende Arbeit ist wie folgt gegliedert. Das nächste Kapitel 1 führt ausführlich in die Problemstellung ein. Es erfolgt dabei eine Klassifikation von Methoden zur Selektivitätsschätzung in Datenbanksystemen. Bisherige Arbeiten auf diesem Gebiet werden vorgestellt und der Rahmen der weiteren Arbeit wird skizziert. In Kapitel 2 werden die Grundlagen sowohl aus dem Bereich der Datenbanksysteme als auch aus dem Bereich der mathematischen Statistik vorgestellt und in Beziehung gesetzt. Substantielle Begriffe wie Selektivität und Selektivitätsschätzung werden mit entsprechenden Begriffen der Statistik definiert. Des weiteren werden wichtige Fehlermaße eingeführt und grundlegende Methoden der nicht-parametrischen Statistik beschrieben. Kapitel 3 stellt die verschiedenen Selektivitätsschätzer vor. Dazu gehören die Histogrammselektivitätsschätzer, für die in Kapitel 3.2 eine Definition sowohl für den univariaten als auch den multivariaten Fall gegeben wird. Dabei werden verschiedene, teilweise neue Partitionierungsverfahren wie z.B. der KD-Baum-Histogrammschätzer oder die Selektivitätsschätzung mit Voronoi-Regionen vorgestellt. In Kapitel 3.4 wird als neue Methode der Kernselektivitätsschätzer eingeführt. Die Behandlung von Randproblemen wird dabei berücksichtigt. Kapitel 3.5 enthält den oben angesprochenen Hybridselektivitätsschätzer zur Lösung von Problemen, die bei nicht-glatten realen Daten auftreten können. Das sowohl bei Histogramm- als auch Kernschätzern auftretende Problem der Wahl des Bandbreitenparameters wird in Kapitel 4 aufgegriffen. Verschiedene Methode zur Wahl einer Bandbreite werden vorgestellt und diskutiert. Die theoretischen Ergebnisse werden in Kapitel 5 mittels Experimenten an künstlichen und realen Daten evaluiert. Die praktischen Ergebnisse zeigen einige interessante Phänomene auf und ergänzen das Verständnis der verschiedenen nicht-parametrischen Verfahren zur Selektivitätsschätzung. Im abschließenden Kapitel 6 erfolgt eine Zusammenfassung der vorliegenden Arbeit und eine Bewertung der theoretischen und praktischen Ergebnisse. Ein Ausblick soll Anreiz für weitere Forschungen auf diesem Gebiet geben.

Inhaltsverzeichnis

1. Einführung

1.1 Problemstellung

Große Datenbestände werden heutzutage üblicherweise in sogenannten *Datenbanksystemen* (*DBS*) abgelegt. Die gespeicherten Daten werden dabei in *Datenbanken* (*DB*) organisiert, während sich das *Datenbankmanagementsystem* (*DBMS*) um Aufgaben wie die Konsistenzerhaltung der Datenbanken sowie Kontrolle und Steuerung von Modifikationen und Zugriffen kümmert. In dieser Arbeit wird davon ausgegangen, daß es sich hierbei um ein *relationales Datenbanksystem* (*RDBS*) handelt. Relationale Datenbanksysteme haben die alten hierarchischen Datenbanksysteme in der Praxis - abgesehen von sogenannten Legacy Systemen - weitestgehend abgelöst. Die in dieser Arbeit entwickelten Verfahren lassen sich aber auch bei objekt-orientierten oder objekt-relationalen Datenbanksystemen einsetzen. Abbildung 1.1 zeigt den Aufbau eines Datenbanksystems. Mittels einer *Data Definition Language* (*DDL*) wird das Schema der Relationen definiert. Transaktionen wie Anfragen oder Modifikationen werden mittels einer deklarativen *Data Modification Language* (*DML*), i.a. der Abfragesprache *SQL*, an den DML- bzw. SQL-Compiler übergeben. Im Modul *Anfragebearbeitung* wird diese deklarative Anfrage in einen Ausführungsplan überführt, der dann an den *Datenbankmanager* zur Ausführung weitergereicht wird. Die *Dateiverwaltung* enthält neben der eigentlichen Datenbasis noch weitere Dateien wie Indexdateien zum effizienten Zugriff auf die Daten oder Redo-Dateien für Sicherungsstrategien (Backup/Recovery). Die in diesem Zusammenhang wichtigen Begriffe über Datenbanksysteme werden in Kapitel 2 genauer definiert. Eine Einführung in Datenbanksysteme findet sich z.B. in [Kemper & Eickler 97].

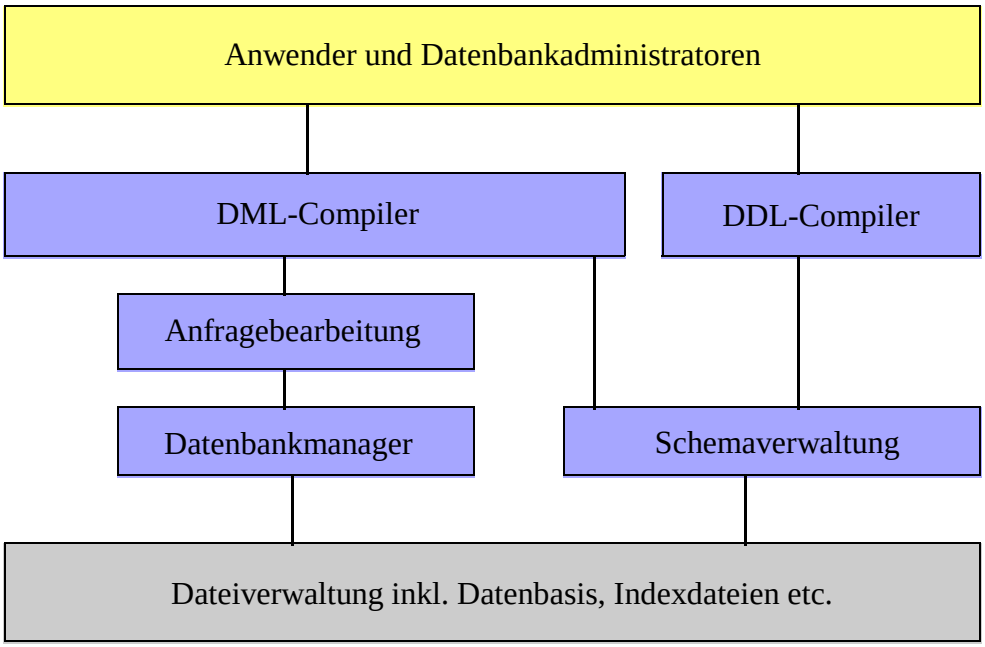


Abbildung 1.1: Aufbau eines Datenbanksystems (stark vereinfacht, angelehnt an [Kemper & Eickler 97])

Eine zentrale Aufgabe eines Datenbankmanagementsystems ist die Transformation von (komplexen) deskriptiven Anfragen, allgemein in der Abfragesprache *SQL* gegeben, in einen effizienten *Ausführungsplan.* Dabei werden die Anfragen zunächst syntaktisch und semantisch analysiert und in äquivalente Ausdrücke der relationalen Algebra transformiert. Die Darstellung ist i.a. nicht eindeutig, da die Möglichkeit besteht auf relationenalgebraischen Ausdrücken Äquivalenzumformungen anzuwenden. Beispielsweise sind Selektionen (s.u.) untereinander vertauschbar und für die Joinoperation (s.u.) gilt die Kommutativität und Assoziativität. Die Auswertung einer Anfrage mittels einer willkürlichen Ausführungsstrategie kann sehr teuer werden, was die Datenbank-Zugriffe und damit zusammenhängend die Ausführungszeit betrifft. Ziel des Anfragebearbeiters ist es durch einen (nahezu) optimalen Ausführungsplan diese Kosten möglichst gering zu halten, im besten Fall sogar den am wenigsten teuren Ausführungsweg zu finden. Hierzu sind geeignete Methoden zur Schätzung der Selektivität, d.h. der erwarteten Antwortmenge, der einzelnen Operatoren erforderlich. Der Ausführungsplan kann durch eine Baumstruktur repräsentiert werden, dessen Knoten die Operatoren einer Algebra darstellen. Selektivitätsschätzung geschieht innerhalb des Operatorbaums von den Blättern ausgehend zur Wurzel (bottom-up), da die Selektivität eines Operators in einem Knoten von der Selektivität der Operatoren in seinen Kinderknoten abhängt. Aufgrund des Problems der Fehlerpropagierung ([Ioannidis & Christodoulakis 91]) ist insbesondere in den Knoten der niedrigeren Ebenen und in den Blättern eine gute Selektivitätsschätzung zu gewährleisten. Wegen der algebraischen Optimierung gehören zu den Knoten in den niedrigeren Ebenen und den Blättern häufig sogenannte Selektionsanfragen. Ergebnis einer solchen Anfrage sind diejenigen Tupel einer Relation, die eine gewisse Selektionsbedingung erfüllen. Da Selektionsanfragen somit zu den wichtigsten und häufigsten Anfragetypen gehören, beschränkt sich diese Arbeit auf die Selektivitätsschätzung von Selektionsanfragen, insbesondere Bereichsanfragen, d.h. solchen Selektionsanfragen bei denen die Attribute in einem angegebenen Bereich liegen müssen. Die in diesem Zusammenhang auftauchenden Begriffe werden in Kapitel 2 genauer definiert.

Abbildung 1.2 zeigt als Beispiel die Baumstruktur des Ausführungsplans der folgenden kanonischen SQL-Anweisung:

select A_1, ..., A_n **from** R_1, ..., R_k **where** P;

Dabei bedeutet σ_P eine Selektionsanfrage mit Bedingung P (hier auf dem Kreuzprodukt der Relationen R_1, ..., R_k) und $\Pi_{A_1, \ldots, A_n}$ die Projektion auf die Attribute A_1, ..., A_n.

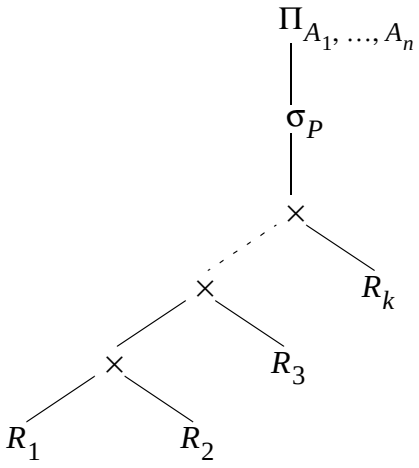


Abbildung 1.2: Beispiel eines kanonischen Ausführungsplans in Baumstruktur

Ursprünglich wurde beim Problem der Selektivitätsschätzung in relationalen Datenbanksystemen von Anfragen auf einem einzigen Attribut einer Relation ausgegangen. In den letzten Jahren ist bei Datenbanksystemen eine deutliche Zunahme von Anwendungen zu beobachten, bei denen mehrdimensionale Indexstrukturen erforderlich sind. Mehrere Attribute werden dabei zusammengefaßt. Beispiele hierfür sind Geo-Datenbanksysteme, aber auch bei der effizienten Speicherung und Abfrage von Multimedia- oder Textobjekten spielen diese eine bedeutende Rolle ([Chiueh 94], [Faloutsos et al. 94], [Hellerstein & Pfeffer 94], [Zezula et al. 96]). Hierfür werden entsprechende Methoden zur Schätzung der Selektivität einer Anfrage gesucht, bei der mehrere Attribute einer Relationen in die Anfragebedingung einbezogen werden. In dieser Arbeit werden die vorgestellten Verfahren zur Selektivitätsschätzung erweitert auf mehrdimensionale Selektionsanfragen.

Bisher wurde davon ausgegangen, daß sich die Anfragen auf Punktobjekte, d.h. reellwertige Vektoren, beziehen. Geoinformationssysteme enthalten aber in der Regel auch räumlich ausgedehnte Objekte [Samet 90]. Eine Bereichsanfrage bzgl. eines räumlich ausgedehnten Objektes wird dabei dahingehend verallgemeinert, daß gefragt wird, welche Objekte von der Bereichsanfrage geschnitten werden. Entsprechend bedeutet die Selektivität einer Bereichsanfrage die Anzahl derjenigen Objekte, die vom Anfragebereich geschnitten werden oder ganz in diesem enthalten sind. Bereichsanfragen werden in dieser Arbeit nicht behandelt, vgl. dazu stattdessen [Schneider 97].

Die hier behandelten Selektivitätsschätzungen für Anfragen sind zu unterscheiden von Kostenschätzungen, die die Zugriffe auf Speicherblöcke mittels geeigneter Indexstrukturen berücksichtigen. Arbeiten hierzu finden sich in ([Belussi & Faloutsos 95], [Pagel et al. 93], [Theodoridis & Sellis 96]). Nichtsdestotrotz können Ergebnisse dieser Arbeit als Grundlage für solche Kostenschätzungen verwendet werden.

1.2 Klassifikation der Verfahren

In diesem Abschnitt wird eine zweidimensionale Klassifikation der Verfahren zur Selektivitätsschätzung vorgeschlagen. Die eine Dimension richtet sich an die Ausprägung der den zu untersuchenden Daten zugrundeliegenden Attribute (Merkmale). Hier wird wie in der Statistik üblich zwischen *quantitativen* und *qualitativen*, zwischen *stetigen* und *diskreten* sowie zwischen *nominalen*, *ordinalen* und *metrischen* Attributen unterschieden. Attribute der Nominalskala unterliegen keiner Reihenfolge und sind nicht vergleichbar, Attribute einer Ordinalskala unterscheiden sich in der Intensität und unterliegen einer Rangfolge und auf Attributen einer metrischen Skala lassen sich zudem Abstände zwischen den Werten interpretieren.

Die zweite Dimension klassifiziert die zugrundeliegenden Verfahren mit Kategorien der mathematischen Statistik ([Büning & Trenkler 94], [Gasser et al. 93]). Hier wird im wesentlichen zwischen *parametrischen* und *nicht-parametrischen* Dichteschätzern unterschieden. Durch die Nähe der Selektivitätsschätzer zur Dichteschätzung lassen sich die verschiedenen Verfahren zur Selektivitätsschätzung ebenfalls in zwei Klassen einteilen:

Parametrische Verfahren

Um die Selektivität einer Anfrage zu schätzen wird bei dieser Klasse der Verfahren davon ausgegangen, daß die wahre Verteilung der Daten zu einer bestimmten angenommenen Verteilungsfamilie gehört. Der Selektivitätsschätzer braucht dann nur noch die freien Parameter dieser Modellfunktion zu schätzen, um die aktuelle Verteilung der Daten eindeutig beschreiben zu können. Rechnet man die zugrunde liegende Verteilung zum Beispiel der Familie der Normalverteilungen zu, so brauchen lediglich der Mittelwert und die Standardabweichung anhand der Daten geschätzt zu werden. Ein weiteres triviales Beispiel ist die Annahme einer Gleichverteilung, wie sie in dem Datenbanksystem *System R* tatsächlich eingesetzt wurde [Selinger et al. 79]. Obwohl der Berechnungsaufwand äußerst gering ist, ist dies i.a. kein praktikables Verfahren. Die wahre Verteilung einer Datenbank ist i.a. nicht bekannt, und eine falsche Modellannahme kann zu extrem schlechten Schätzungen führen.

Nichtparametrische Verfahren

Verfahren dieser Klasse gehen nicht von der Annahme aus, daß die den Daten zugrunde liegende Verteilung zu einer speziellen Verteilungsfamilie gehört. Stattdessen wird die Selektivität auf Basis einer Stichprobe aus der Datenbank geschätzt (vgl. Abbildung 1.3). Beispiele solcher Verfahren sind die direkte Selektivitätsschätzung aus der Stichprobe, die Selektivitätsschätzung mittels Histogrammschätzer oder die Selektivitätsschätzung mittels Kernschätzer, wie in dieser Arbeit vorgeschlagen. (vgl. Abbildung 1.4).

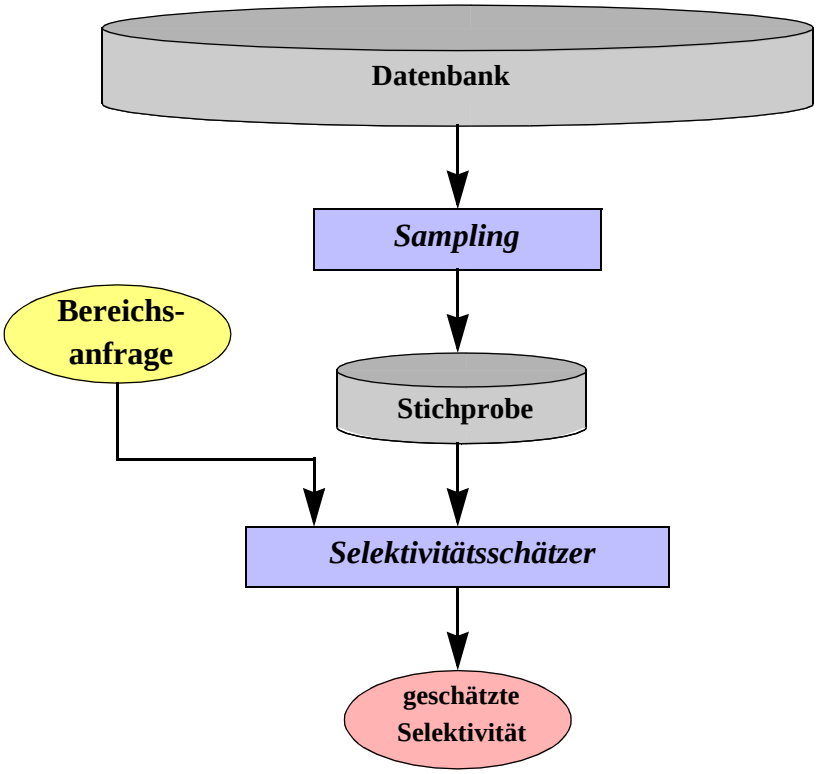


Abbildung 1.3: Generelles Verfahren zur nicht-parametrischen Selektivitätsschätzung

Eine weitere Alternative nicht-parametrischer Selektivitätsschätzung besteht in der Schätzung mittels Funktionenapproximation. Dazu gehören zum einen Orthogonalreihenschätzer unter der Verwendung der Fourierentwicklung oder von Wavelets als auch die Verwendung von Spline-Funktionen oder Polynomen. Diese werden bestimmt mittels Minimierung des kleinsten quadratischen Fehlers. Ein diesem Vorgehen ähnlicher Ansatz zur Selektivitätsschätzung ist in [Chen & Roussopoulos 94] beschrieben. Der Nachteil dieser Methode ist, daß ein genaue Schätzung einen hohen Grad der Polynome erfordert. Ein hoher Grad der Polynome führt jedoch oft zu dem Problem, daß Oszillationen und Rundungsfehler auftauchen können. Weiter unten wird auf die Verwandschaft dieser Verfahren mit Kernschätzern hingewiesen.

Die hier vorgestellte Einteilung in parametrische und nicht-parametrische Verfahren deckt sich weitestgehend mit der in der Literatur über Selektivitätsschätzer vorgestellten. In [Ioannidis & Poosala 95] werden die Schätzverfahren mittels Zufallsstichprobe einer extra Klasse zugerechnet. Nicht-parametrische Verfahren werden unterteilt in Histogramm-Verfahren und algebraische Verfahren, wobei zu letzteren Schätzverfahren mittels Polynomapproximation gezählt werden. Diese Polynomapproximationen werden in [Poosala et al. 96] fälschlicherweise zu den parametrischen Verfahren gezählt. [Chen & Roussopoulos 94] folgt der Einteilung von Poosala und Ioannidis, ordnet aber die Schätzverfahren mittels Polynomapproximationen einer weiteren separaten Klasse zu.

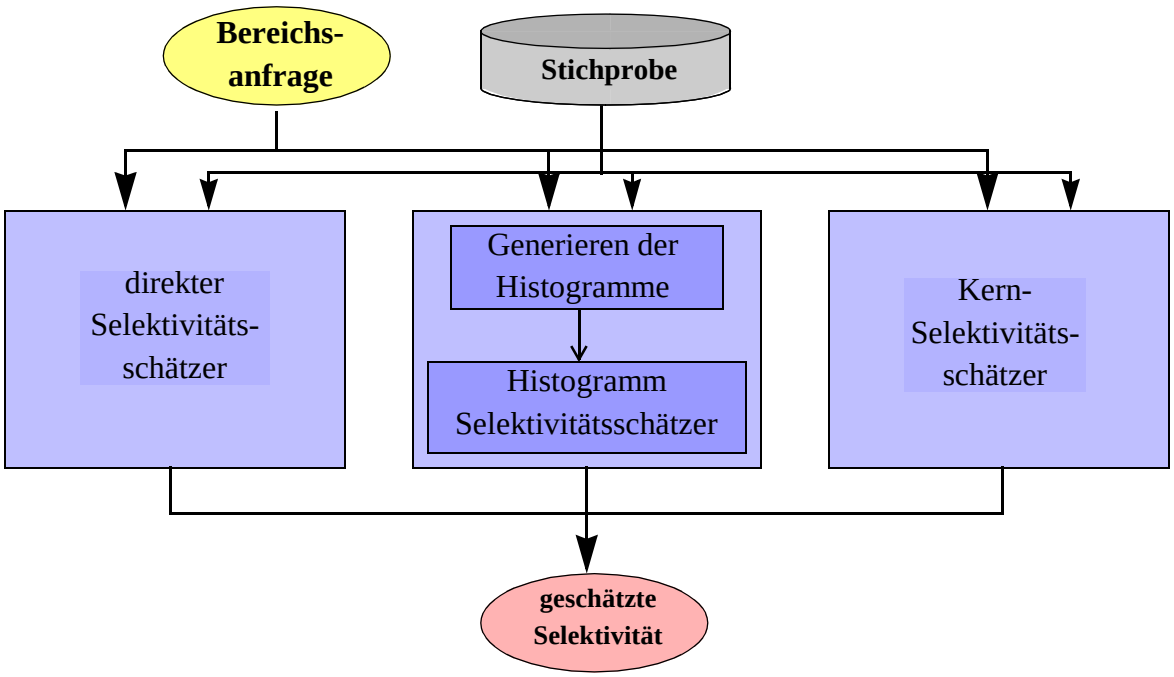


Abbildung 1.4: Verschiedene Verfahren zur nicht-parametrischen Selektivitätsschätzung

Für alle hier vorgestellten nicht-parametrischen Verfahren wird das Vorliegen einer Zufallsstichprobe vorausgesetzt (vgl. Abbildung 1.4). Eine im Bereich Datenbanksysteme sehr weit verbreitete Methode zum Ziehen einer Zufallsstichprobe aus einer Datenbankinstanz ist das Reservoir Sampling von Vitter [Vitter 85]. Es besitzt den Vorteil, daß die Größe der Instanz nicht im voraus bekannt sein muß.

Das direkte Schätzen der Selektivität anhand einer Zufallsstichprobe wird in der Datenbank-Literatur auch oft Selektivitätsschätzung mittels Sampling genannt. Grundlage hierfür ist das Vorliegen einer Zufallsstichprobe (*random sample set*). Es werden dazu lediglich die Elemente der Stichprobe im Anfragebereich gezählt. Im univariaten Fall ist diese direkte Selektivitätsschätzung äquivalent zur Selektivitätsschätzung mittels empirischer Verteilungsfunktion, vgl. hierzu Kapitel 3.1. Um jedoch die gewünschte Genauigkeit zu erreichen, ist i.a. eine sehr große Stichprobe erforderlich (Konvergenz der Ordnung $O(n^{1/2})$, siehe Kapitel 4.5). Von daher sind Selektivitätsschätzer von Interesse, die i.a. mit einer geringeren Stichprobe auskommen, um die gleiche Genauigkeit zu erreichen.

Bei *Histogrammschätzern* wird zunächst der Datenraum in disjunkte Partitionen unterteilt, die die Domäne vollständig überdecken. Die Partitionen werden auch *Histogrammklassen* oder *Histogrammbins* genannt. Für jedes Histogrammbin wird nun die Anzahl der Stichprobenwerte bestimmt, die in diesem Bin liegen. Zur Schätzung der (absoluten / relativen) Selektivität werden nun die Schnittmengen aus Anfragebereich und Histogrammbins gebildet und mit der (absoluten / relativen) Anzahl der Elemente eines Histogrammbins gewichtet, siehe Kapitel 3.2.

Die Ordnung der Konvergenz bei Histogrammschätzern (mit konstanter Binweite) ist $O(n^{2/3})$ (vgl. Kapitel 4.5) und damit deutlich besser als bei der direkten Selektivitätsschätzung aus der Stichprobe.

In [Blohsfeld 98] wurde der *Kernselektivitätsschätzer* eingeführt, ein weiteres nicht-parametrisches Verfahren zur Selektivitätsschätzung von Selektionsanfragen, insbesondere solchen mit einem Bereichsprädikat. Kerndichteschäzer haben in den letzten Jahren immer mehr Aufmerksamkeit auf dem Gebiet der mathematischen Statistik erlangt, vgl. [Büning & Trenkler 94], [Rao 93], [Müller 88], [Silverman 86], [Wand & Jones 95]. Bisher wurden solche Methoden jedoch nicht zur Selektivitätsschätzung in Datenbanksystemen verwendet. Auch hier wird analog zum Histogrammselektivitätsschätzer das Vorliegen einer Zufallsstichprobe vorausgesetzt. Das Grundprinzip der Kerndichteschätzung besteht darin, die Dichtefunktion durch geeignete Überlagerung verschiedener parametrisierter Kernfunktionen zu schätzen. Lage und Gestalt der Kernfunktionen hängen dabei von den Elementen der Zufallsstichprobe ab. Die Weite einer Kernfunktion wird durch einen Glättungsparameter, die sogenannte Bandbreite, festgelegt, die gleichzeitig die Glattheit der Schätzfunktion bestimmt. Am Rand des Wertebereich ist die Dichtefunktion i.a. unstetig, so daß hier spezielle Randbehandlungsmethoden erforderlich sind, siehe Kapitel 3.4. Bei den Kernschätzern wird eine Ordnung der Konvergenz von $O(n^{4/5})$ erreicht (siehe Kapitel 4.5). Dies ist wesentlich besser als bei den beiden zuvor beschriebenen nicht-parametrischen Verfahren und kommt der optimalen Konvergenzordnung von $O(n)$ schon recht nahe.

Die Wahl des Glättungsparameters ist sowohl bei Histogramm- als auch bei Kernschätzern von großer Bedeutung. In der mathematischen Statistik sind verschiedene Verfahren zur Bestimmung einer Bandbreite bekannt, die den asymptotischen durchschnittlichen integrierten quadratischen Fehler (engl. *asymptotic mean integrated squared error*, kurz *AMISE*) (möglichst) minimiert. Da die asymptotisch optimale Bandbreite i.a. abhängig von der wahren Dichte ist, existieren verschiedene Verfahren um diese zu schätzen. Diese Verfahren werden in Kapitel 4 vorgestellt und diskutiert.

1.3 Bisherige Arbeiten

In diesem Abschnitt sollen bisherige Arbeiten zur Selektivitätsschätzung von Bereichsanfragen in Datenbanksystemen beleuchtet werden. Obwohl in den meisten Fällen die statistische Zuordnung nicht explizit geleistet wurde, lassen sich diese Arbeiten vor dem statistischen Hintergrund einordnen. In den meisten Arbeiten handelt es sich dabei um nicht-parametrische Methoden zur ad hoc Selektivitätsschätzung.

1.3.1 Parametrische Methoden zur Selektivitätsschätzung

Da in dieser Arbeit nur nicht-parametrische Methoden verglichen werden, sollen die parametrischen Methoden zur Selektivitätsschätzung nur kurz erwähnt werden. Bei parametrischen Methoden geht man davon aus, daß die den Daten zugrunde liegende Verteilung zu einer

bekannten Verteilungsfamilie gehört, so daß lediglich die unbekannten Parameter dieser Verteilung zu schätzen sind. Dies wären z.B. bei der Normalverteilung $N(\xi,s)$ der Mittelwert ξ und die Standardabweichung s (bzw. die Varianz s^2).

Die einfachste Verteilungsannahme wäre die *Annahme der Gleichverteilung* der Daten auf dem gesamten Wertebereich. Hierzu braucht lediglich die Anzahl der Elemente der Datenbank-Instanz und die Domäne bekannt zu sein. Dieses Verfahren ist eines der ersten Verfahren, welches in einem Datenbanksystem - dem System-R [Selinger et al. 79] - zur Selektivitätsschätzung eingesetzt wurde. Aus diesem Grunde wurde es bei unseren Experimenten als Referenzwert berücksichtigt. Allerdings ist davon auszugehen, daß reale Daten i.a. nicht gleichverteilt sind, so daß hier entsprechend schlechte Ergebnisse zu erwarten sind. Wie schlecht die Annahme der Gleichverteilung bei Daten ist, die nicht gleichverteilt sind, zeigen die Experimente bei normalverteilten und exponentialverteilten Daten in Kapitel 5.

Weitere Arbeiten mit anderen Annahmen der zugrundeliegenden Verteilungsfamilie sind in [Chen & Roussopoulos 94] zitiert, werden hier aber nicht näher beleuchtet.

1.3.2 Einfache nicht-parametrische Methoden zur Selektivitätsschätzung

Eine der einfachsten (nicht-parametrischen) Methoden zur Selektivitätsschätzung ist der Spezialfall eines Histogrammschätzers (s.u.) mit nur einem einzigen Bin. Da beim Histogrammschätzer pro Bin die Gleichverteilungsannahme gilt, ist diese Methode äquivalent zu der im vorigen Abschnitt erwähnten Methode der parametrischen Dichteschätzung auf Basis einer Gleichverteilungsannahme der Daten.

Ein weiteres häufig in Datenbanksystemen eingesetztes nicht-parametrisches Verfahren zur Selektivitätsschätzung ist die *direkte Selektivitätsschätzung* anhand der Zufallsstichprobe. In der Literatur findet man in diesem Zusammenhang oft den Begriff „Sampling". Im univariaten Fall entspricht dies der Selektivitätsschätzung mittels empirischer Verteilungsfunktion. In einer vom Autor betreuten Diplomarbeit [Schneider 97] wurde die Selektivitätsschätzung mittels empirischer Verteilungsfunktion mit Methoden der mathematischen Statistik untersucht. Weitere Beispiele zur Selektivitätsschätzung anhand einer Zufallsstichprobe im Bereich Datenbanksysteme finden sich in [Lipton & Naughton 90] und [Haas & Swami 92], für eine generelle Diskussion von Stichprobenverfahren vergleiche auch die umfassende Arbeit von [Olkem & Rotem 94]. Aufgrund der Einfachheit und weiten Verbreitung dieser Methode zur Selektivitätsschätzung wird sie in den Experimenten berücksichtigt und die Ergebnisse dienen als Referenzwerte für die anderen Selektivitätsschätzer.

1.3.3 Histogramme zur Selektivitätsschätzung

Histogramme zur Selektivitätsschätzung wurden in zahlreichen Arbeiten präsentiert. Bisher fehlte es jedoch an einer expliziten Definition eines Histogrammselektivitätsschätzers. Diese wird in dieser Arbeit in Abschnitt 3.2 vorgenommen. Lediglich in [Poosala et al. 96] findet sich der Versuch einer Definition mittels einer "Konstruktionsmethode" für univariate Histogramme, die in [Poosala & Ioannidis 97] auch auf multivariate Histogramme erweitert wurde.

Im Unterschied zur vorliegenden Arbeit betrachten die Autoren hingegen keine großen und stetigen sondern kleine und diskrete Domänen. Dabei wird in [Poosala et al. 96] teilweise davon ausgegangen, daß in der Stichprobe sämtliche möglichen Werte einer Relation enthalten sind. Um dies zu gewährleisten, muß gegebenenfalls die gesamte Datenbank-Instanz als Stichprobe herangezogen werden. Dies hat natürlich ein wesentlich höheren Speicheraufwand zur Folge, so daß dieses Verfahren für realistische Anwendungen mit großen Datenbanken nicht praktikabel ist. Mögliche Werte, die nicht in der Stichprobe (bzw. der Instanz) liegen, bekommen die Häufigkeit 0 zugewiesen. [Poosala et al. 96] empfehlen die Stichprobenwerte entsprechend ihrer Häufigkeit absteigend zu sortieren, so daß sich ihrer Meinung nach in vielen Fällen eine der Zipf-Verteilung ([Zipf 49]) ähnliche Verteilung dieser Häufigkeiten ergibt. Die Experimente bei [Poosala et al. 96] beziehen sich somit auch auf künstliche Zipf-Verteilungen der möglichen Werte mit kleinem (diskreten) Wertebereich und ebenfalls einer Zipf-Verteilung in den Häufigkeiten. Die Ergebnisse sind daher nicht übertragbar auf die in dieser Arbeit vorausgesetzten stetigen Verteilungen mit hoher Kardinalität der Domäne.

Bekannte zur Selektivitätsschätzung verwandte Methoden im univariaten Fall sind das *Equi-Width*- und das *Equi-Depth-Histogramm*. Diese Methoden werden von [Poosala et al. 96] um weitere univariate Verfahren zur Partitionierung erweitert. Aufgrund der dortigen Voraussetzungen an die Daten lassen sich die Verfahren nur teilweise auf die in dieser Arbeit getroffen Voraussetzungen übertragen (siehe oben). Die Autoren berichten, daß in ihren Experimenten das *Max-Diff-Histogramm* die besten Resultate geliefert hat. Es wird daher in der vorliegenden Arbeit ebenfalls berücksichtigt.

Ein weiteres verwandtes Verfahren besteht in der Verwendung von *Average Shifted Histogrammen* [Scott 92], welches in dieser Arbeit erstmalig zur Selektivitätsschätzung angewendet wird. Vorteil ist insbesondere, daß dabei die starke Abhängigkeit der Histogrammschätzer vom Startpunkt vermieden wird. Auch dieses Verfahren wird zur Anwendung auf multivariate Daten erweitert.

Ein möglicher Ansatz, die für univariate Daten entwickelten Histogrammschätzer auch bei multivariaten Daten zu benutzen, besteht darin, die (diskretisierten) multivariaten Daten mittels einer Ordnung auf eine (diskrete) eindimensionale Domäne abzubilden. Auf dieser können dann prinzipiell die oben beschriebenen univariaten Schätzverfahren angewendet werden. Solche Verfahren zur Abbildung mehrdimensionaler auf eindimensionale Räume wurden bereits zuvor für effiziente Zugriffe auf mehrdimensionale und räumliche Daten mittels eines zweiten Schlüssels angewendet [Faloutsos & Roseman 89]. Hierbei wurde als raumfüllende Kurve die Hilbert-Kurve verwendet.

In der Diplomarbeit von [Buskamp 97] wurde die multivariate Domäne mittels einer *raumfüllenden Kurve* auf eine eindimensionale Domäne abgebildet. Dazu wurden die multivariaten Daten zuvor auf ganzzahlige Attribute transformiert. In der Diplomarbeit wurde als raumfüllende Kurve die Z-Kurve (Lesbegue-Kurve) verwendet, aber auch andere raumfüllende Kurven wie z.B. die Hilbert-Kurve sind möglich [Poosala & Ioannidis 97]. Auf die mittels der raumfüllenden Kurve abgebildeten univariaten Daten können nun die bisher bekannten univariaten nichtparametrischen Selektivitätsschätzer angewendet werden. In der Diplomarbeit wurden

dazu Histogrammschätzer verwendet. Zu beachten ist dabei, daß ein im mehrdimensionalen Ursprungsraum zusammenhängendes (mehrdimensionales) Anfrageintervall bei der Abbildung mittels einer raumfüllenden Kurve durchaus auf mehrere nicht zusammenhängende eindimensionale Intervalle abgebildet werden kann. Die Selektivitätsschätzung im eindimensionalen erfolgt dann zunächst für diese nichtzusammenhängenden Intervalle getrennt und wird anschließend aufsummiert. Da die entstehenden eindimensionalen Intervalle dabei i.a. von geringer Größe sind (teilweise nur mit 1 bis 4 Elementen), dafür aber äußerst zahlreich auftreten, kann der Rechenaufwand zur Berechnung sehr groß werden, so daß dieses Verfahren für realistische Anwendungen bei großen Datenbanken nicht praktikabel ist. Die Verwendung von raumfüllenden Kurven zur Selektivitätsschätzung und die dabei auftretenden Berechnungsprobleme werden ausführlich in Kapitel 3.2.2 besprochen.

Weitere bekannte Verfahren zur multivariaten Selektivitätsschätzung mit Histogrammschätzern partitionieren direkt die multivariate Domäne zur Bildung multivariater Histogramme.

Ein früher Ansatz zur Selektivitätsschätzung multivariater Daten findet sich in [Muralikrishna & DeWitt 88]. Von den Autoren wird ein Ansatz vorgeschlagen, Equi-Depth Histogramme auf den multivariaten Fall auszudehnen und anschließend zur Selektivitätsschätzung zu verwenden. Dieses Problem besitzt allerdings im multivariaten Fall keine eindeutige Lösung. In dem von den Autoren vorgeschlagenen Verfahren wird zunächst eine Achse (Dimension) willkürlich ausgewählt. Der Datenraum wird nun durch $k_1 - 1$ bzgl. dieser Achse orthogonale Hyperebenen so partitioniert, daß in allen k_1 Partitionen gleich viele Elemente (N / k_1) liegen. Nun wird - wieder willkürlich - eine weitere Achse gewählt, so daß nun jede der vorherigen Partitionen durch $k_2 - 1$ bzgl. dieser zweiten Achse orthogonale Hyperebenen weiter unterteilt wird, so daß nun in diesen insgesamt $k_1 \cdot k_2$ Teilpartitionen wieder jeweils gleich viele Elemente ($N / (k_1 \cdot k_2)$) liegen. Dieses Verfahren wird so lange fortgesetzt bis der Datenraum bzgl. allen d Achsen partitioniert wurde, so daß jede der $\prod_{i=1}^{d} k_i$ endgültigen Partitionen $N / \prod_{i=1}^{d} k_i$ Elemente enthält. Dabei wird hier davon ausgegangen, daß N durch $\prod_{i=1}^{d} k_i$ ganzzahlig teilbar ist. Ansonsten können sich leichte Abweichungen bzgl. der Anzahl der Elemente pro Partition ergeben. [Muralikrishna & DeWitt 88] vergleichen ihre Methode mit den bekannten multivariaten Equi-Width Histogrammen zur Selektivitätsschätzung, bei denen das Volumen der einzelnen Partitionen (Bins) für alle Bins gleich ist. Dazu verwenden sie gleichverteilte, normalverteilte und Zipf-verteilte künstliche Testdaten mit kleiner Domäne (Wertebereich ganzzahlig von 1 bis 240) sowie sehr große Anfragebereiche. Reale Testdaten werden nicht untersucht. Die Ergebnisse des Equi-Width- und des Equi-Depth-Histogrammschätzers sind bei gleich- und bei normalverteilten Testdaten gleichwertig, während bei schiefen Zipfverteilten Datensätzen der ED-Histogrammschätzer einen deutlich geringeren Fehler aufweist. Für die optimale Wahl der Anzahl der Histogrammbins gibt es [Muralikrishna & DeWitt 88] keinerlei Hinweise - in den Experimenten wählen die Autoren willkürlich die Aufteilung 5x5, 10x10 und 20x20 Bins unabhängig von der wahren Verteilung. Aufgrund der unterschiedlichen Testvor-

aussetzungen (vgl. Kap. 5.3) werden das Equi-Width und das Equi-Depth-Histogramm nach der Methode von [Muralikrishna & DeWitt 88] in dieser Arbeit bei den Experimenten zur bivariaten Selektivitätsschätzung berücksichtigt (siehe Kap. 5.6).

Weitere Verfahren zur multivariaten Selektivitätsschätzung mit Histogrammen verwenden zur Partitionierung aus multidimensionalen Datenbanken bekannte *Indexstrukturen.* Voraussetzung ist, daß bei der Erzeugung der Indexstruktur eine vollständige und disjunkte Partitionierung der Domäne entsteht. Anstelle der Daten ist dabei in den Knoten lediglich die Anzahl der Daten bzw. Stichprobenelemente in diesem Bereich gespeichert. Ein Vorteil dieser Methoden liegt sicherlich in dem schnellen Zugriff über die Indexstruktur auf die Information der Histogramm-Bins. Bisher bekannte Verfahren zur Selektivitätsschätzung beruhen auf dem Multi-Level-Grid-File ([Whang et al. 94]) und dem KD-Baum ([Buskamp 97], [Poosala & Ioannidis 97]). Die Methoden werden im folgenden kurz beschrieben.

[Whang et al. 94] verwenden als mehrdimensionale Indexstruktur zur Partitionierung das *Multi-Level-Grid-File.* Durch Hinzufügen von Datenpunkten wird sukzessive eine Indexstruktur aufgebaut. Ausgangspartition ist die gesamte Domäne. Wird bei der Anzahl der Elemente einer Partition ein gewisser Schwellwert überschritten, so wird die Partition in zwei neue gleich große Bereiche geteilt. Ziel ist es dabei eine Partitionierung zu erhalten, bei der in jeder Partition ungefähr gleich viele Elemente enthalten sind. Insofern besteht hierbei eine gewisse Ähnlichkeit zu Equi-Width-Histogrammen. Allerdings wird bei dem Multi-Level-Grid-File eine hierarchische Partitionierung aufgebaut, wobei man bei der Anwendung zur Selektivitätsschätzung bei jeder Partition die Anzahl der darin enthaltenen Datenpunkte bzw. Stichprobenelemente zusätzlich abspeichert. Die unterschiedlichen Ebenen (levels) der Hierarchie repräsentieren somit unterschiedliche Histogramme zur zugrundeliegenden Dichte. Auch hier erfolgt ein schneller Zugriff auf die einzelnen Histogrammbins aufgrund der Partitionierung mittels einer Indexstruktur. [Whang et al. 94] bauen in ihren Experimenten lediglich Histogramme auf der gesamten Datenbasis und nicht auf Stichproben auf. Als Ergebnis berichten sie, daß bei tieferen Ebenen (mehr Histogrammbins) die Qualität der Selektivitätsschätzung besser wird. Eine Betrachtung der optimalen Anzahl der Histogrammbins erfolgt also bei den Autoren nicht. In [Whang et al. 94] werden ausführliche Experimente beschrieben bei gleich-, normal- und exponentialverteilten künstlichen Testdaten mit sehr großer ganzzahliger Domäne, aber nicht mit realen Daten. Die Anfragegröße liegt zwischen 20% und 0,1% der Domänengröße. Die Autoren variieren die Verteilung, die Größe der Anfragen, sowie die Feinheit der Partitionierung (durch Wahl unterschiedlicher Ebenen), vergleichen ihre Ergebnisse aber nicht mit anderen Histogrammverfahren zur Selektivitätsschätzung wie dem Equi-Width- oder dem Equi-Depth-Histogrammschätzer.

Eine als Indexstruktur bzgl. der Partitionierung noch größere Flexibilität gegenüber dem Multi-Level-Grid-File bietet der *KD-Baum* (s. [Samet 90]), der daher in der Diplomarbeit von [Buskamp 97] als zugrunde liegende Indexstruktur gewählt wurde. Dieses Verfahren führt zu einer Partitionierung, die zwar immer noch rechteckige Bins erzeugt, diese sich aber in Lage und Größe besser der Struktur der Daten anpassen. Insbesondere können die Partitionen an beliebiger Stelle geteilt werden (nach vorgegebenen Kriterien), so daß die beiden neuen Partitionen

unterschiedlich groß sein können. Vorteil ist weiterhin der schnelle Zugriffe auf die einzelnen Bins aufgrund der Baum-Struktur des Verfahrens. Aufgrund der hierarchischen Struktur ist Ziel eine Partitionierung unterschiedlicher Feinheit zu erreichen, so daß die Güte der Selektivitätsschätzung dieser Feinheit angepaßt werden kann. Histogrammselektivitätsschätzer auf der Basis von KD-Bäumen werden ausführlich in dieser Arbeit untersucht und in Kapitel 3.2.3 beschrieben.

In [Poosala & Ioannidis 97] finden sich weitere multivariate Histogrammselektivitätsschätzer. Die Autoren beschränken sich in der Veröffentlichung auf eine sehr kleine und diskrete Domäne mit einer hohen Dichte (in ihren Experimenten 1000000 bis 5000000 Tupel, davon nur 50 bis 200 verschiedene Attributwerte je Dimension), aber einige der Verfahren lassen sich auf große und stetige Domänen übertragen. Dazu gehört der Algorithmus MHIST-p, bei der die Domäne ähnlich wie beim oben beschriebenen KD-Baum-Verfahren partitioniert wird. Der Wert p bestimmt dabei, in wieviele Teilpartitionen eine betrachtete Partition weiter unterteilt wird, so daß sich nur für p = 2 ein direkter Vergleich mit dem KD-Baum Selektivitätsschätzer ergibt. Allerdings berichten die Autoren auch, daß sie mit p = 2 die besten Ergebnisse erzielt hätten. Die Splitstrategie in [Poosala & Ioannidis 97] ist etwas eingeschränkt im Vergleich zum KD-Baum Selektivitätsschätzer. Es werden von den Autoren lediglich Randverteilungen der Partitionen betrachtet und anhand von Verfahren bei univariaten Histogrammselektivitätsschäter bewertet ([Poosala et al. 96], siehe auch oben). Diese lassen sich nur teilweise auf große stetige Domänen übertragen (siehe oben). Auf die Splitstrategie wird etwas genauer eingegangen im Zusammenhang mit dem KD-Baum Selektivitätsschätzer in Kapitel 3.2.3. Einen Hinweis auf die zu wählende Anzahl Histogrammbins gibt es nicht.

In [Matias et al. 98] wurde ein Verfahren entwickelt, welches Histogrammbins auf Grundlage einer *Wavelet-Partitionierung* erzeugt. Die Autoren beschränken sich dabei auf diskrete Daten, eine Erweiterung auf stetige Daten ist nicht ohne weiteres möglich. Durch sukzessive paarweise Zusammenfassung benachbarter Werte bzw. von deren Häufigkeiten werden Häufigkeitshistogramme auf unterschiedlichen Ebenen generiert. Sei beispielsweise die Anzahl der unterschiedlichen möglichen Werte 16, so werden im univariaten Fall Equi-Width-Histogramme mit 8, 4, 2 und schließlich einem Bin gebildet. Diese werden mittels Wavelet-Koeffizienten in einer Reihe repräsentiert. Unter Benutzung der Haar-Koeffizienten lassen sich die Histogramme rekonstruieren, unter Benutzung von linearen Koeffizienten werden die zugehörigen Häufigkeitspolygone erzeugt. Durch das Abschneiden der Koeffizienten ab einer bestimmten Position wird die Ebene der Partitionierung festgelegt. Dieses Vorgehen läßt sich auch auf multivariate diskrete Daten anwenden. [Matias et al. 98] untersuchen die Güte ihres Verfahrens im uni- und bivariaten Fall. Im univariaten Fall berichten sie die Ergebnisse auf künstlichen Zipfverteilten Testdaten mit sehr kleiner Domäne (4096 mögliche verschiedene Werte). Der Typ der Abfragen variiert, wobei bei einem Testlauf Anfragen der festen Größe $10 / 4096 \approx 0,24\%$ verwendet werden. Die Autoren vergleichen die Ergebnisse mit dem Max-Diff-Histogrammverfahren von [Poosala et al. 96] sowie mit der direkten Selektivitätsschätzung auf einer Stichprobe. Die Anzahl der Bins bzw. die Stichprobengröße wird willkürlich in Abhängigkeit einer Speichergröße (42 reelle Zahlen) gewählt, so daß sich 21 Wavelet-Koeffizienten ergeben (entspricht 14 Bins), 14 Bins beim Max-Diff-Histogramm und 42 Stichprobenelemente. Die Histogramme

wurden auf Grundlage der gesamten Datenbasis und nicht auf der Stichprobe gebildet. Die Autoren berichten hierfür lediglich den absoluten Fehler. Danach schneidet das Equi-Width-Häufigkeitshistogramm (Haar-Wavelet) am schlechtesten ab und das Equi-Width-Häufigkeitspolygon (lineares Wavelet) am besten. Auch im bivariaten Fall untersuchen die Autoren ihr Verfahren anhand von künstlichen Zipfverteilten Testdaten sowie an realen Testdaten des U.S. Census Bureaus ([census data]) und vergleichen ihre Ergebnisse mit denen des MHist-2 Verfahrens von [Poosala & Ioannidis 97] (siehe oben). Die Domäne der künstlichen Testdaten ist sehr klein (256 mögliche verschiedene Werte je Dimension), die Anzahl der Histogrammbins wird wieder willkürlich in Abhängigkeit einer Speichergröße (210 reelle Zahlen) gesetzt, so daß sich 30 Bins für das MHIST-2 Verfahren und 70 Wavelet-Koeffizienten ergeben. Als Anfragen werden Intervalle der Form $[(0,0), (b_1,b_2)]$ gewählt. Dadurch sind die Ergebnisse insbesondere bei Zipfverteilten Daten nicht unbedingt aussagekräftig, vgl. Kapitel 5.3.3. Nach den in [Matias et al. 98] berichteten Ergebnissen schneidet das Verfahren der Autoren besser ab als das MHIST-2 Verfahren. Da es sich bei diesem Ansatz letztendlich um ein Equi-Width-Histogramme handelt, wird die Methode von [Matias et al. 98] nicht weiter betrachtet. Equi-Width-Histogramme zur Selektivitätsschätzung werden explizit in dieser Arbeit in den folgenden Kapiteln behandelt.

Alle bisherigen multivariaten auf Histogramme basierenden verwenden eine rechteckige Partitionierung. Ein in der Diplomarbeit von [Schneider 97] vorgeschlagenes Verfahren zur multivariaten Selektivitätsschätzung beruht auf einer Partitionierung mittels Voronoi-Regionen. Der Datenraum wird dabei so in vorgegebene Voronoi-Regionen partitioniert, daß jedes Element zu den Zentren seiner eigenen Voronoi-Region einen geringeren Abstand bzgl. einer gewissen Metrik hat als zu den Zentren der anderen Regionen. Die Zentren der Voronoi-Region heißen auch Ankerpunkte. Ein zunächst offenes Problem ist die Wahl der Ankerpunkte. In [Schreiber 91] wird ein inkrementelles Verfahren zur Bildung einer k-means ähnlichen Clusterung ([Hartigan 75]) auf Basis von Voronoi-Partitionierungen vorgeschlagen. Dieses Verfahren läßt sich anwenden zur inkrementellen Bildung von Voronoi-Regionen nach vorgegebenen Kriterien. In [Schreiber 91] wird dabei als Kostenfunktion die entsprechende Funktion aus dem k-means Verfahren verwendet zur Erzielung einer k-means artigen Clusterung. Es lassen sich aber auch andere Kriterien verwenden, wie z.B. daß man eine (ungefähr) gleiche Anzahl von Daten (Stichprobenelementen) in allen Voronoi-Regionen erreichen möchte, was einer Equi-Depth-Voronoi Partitionierung entspräche. Die Ankerpunkte werden dabei aus einer Stichprobe als Zentren von Clustern gebildet, welche mittels eines speziellen dem k-means Verfahren ähnlichen Clusterverfahren bestimmt wurden. Dieses Clusterverfahren zur Bestimmung der Ankerpunkte sowie die Histogramm-Selektivitätsschätzung mittels Voronoi-Regionen wird in Kapitel 3.2.4 beschrieben.

1.3.4 Weitere Methoden zur Selektivitätsschätzung

In [Chen & Roussopoulos 94] findet sich ein Verfahren zur Selektivitätsschätzung mittels *Polynomapproximation.* Es handelt sich um ein Verfahren zur Minimierung eines Fehlerterms mittels einer nichtparametrischen Schätzfunktion, das zusätzlich noch um eine Lernkomponente erweitert wurde. Das Verfahren berücksichtigt die korrekten Selektivitäten vorheriger

Anfragen, um die Parameter der Polynome anzupassen. Ein sogenannter Vergessensterm steuert in dem rekursiven Algorithmus die Aktualität der zu berücksichtigen Selektivitäten. Der Nachteil von Verfahren mittels Polynomapproximation ist, daß ein genaue Schätzung einen hohen Grad der Polynome erfordert. Ein hoher Grad der Polynome führt jedoch oft zu dem Problem, daß Oszillationen und Rundungsfehler auftauchen können. Insbesondere sind möglicherweise auftretende negative Werte bei der Dichte- bzw. Verteilungsschätzung unerwünscht. Die Autoren schlagen die Verwendung von Polynomen des Grades 6 vor. Ein weiteres Problem ist laut [Chen & Roussopoulos 94] die Initialisierung des Algorithmus, da die Güte hiervon sehr stark abhängt. Die Autoren schlagen daher eine künstliche Lernphase mit gleichverteilten Daten vor, um die Funktion zunächst an eine Gleichverteilung zu adaptieren. Die Ergebnisse ihres Algorithmus werden [Chen & Roussopoulos 94] mit einem Verfahren mittels Polynomapproximation ohne Lernkomponente verglichen. Bei bestimmten Verteilungen zeigt sich nach einer gewissen Lernphase (100 Iterationen) eine starke Verbesserung, während das Verfahren bei anderen Verteilungen wie z.B. der Normalverteilung leicht schlechter abschneidet.

Ein weiteres nicht-parametrischen Verfahren zur Selektivitätsschätzung wurde in der Diplomarbeit von [Blohsfeld 98] (s.a. [Blohsfeld et al. 99]) vorgestellt. Es handelt sich dabei um ein völlig neues Verfahren, das auf der Verwendung von sogenannten *Kernschätzern* beruht. Diese haben gegenüber Histogrammschätzern den Vorteil, daß sie bessere statistische Eigenschaften besitzen, und daß keine Histogramme apriori erzeugt und gespeichert werden müssen. Gegenüber der Polynomapproximation haben sie den Vorteil, daß keine Rundungsprobleme auftauchen und keine Oszillationen oder negativen Werte entstehen können. Die Verwendung von univariaten und multivariaten Kernschätzern zur Selektivitätsschätzung wird in dieser Arbeit vorgestellt und ausführlich untersucht.

Um die Vorteile von Histogrammschätzer und Kernschätzern zu kombinieren, wird in dieser Arbeit als neues Verfahren der *Hybrid-Selektivitätsschätzer* vorgestellt. Dabei werden Histogrammbins so gewählt, daß ihre Ränder an Sprungstellen der zugrundeliegenden Dichtefunktion liegen. Innerhalb dieser Histogrammbins wird anstelle der Gleichverteilungsannahme ein Kernschätzer auf die Daten angewendet. Die Ergebnisse bzgl. der einzelnen Histogrammbins werden anschließend zur Selektivitätsschätzung der Anfrage kombiniert. Mehr noch als bei Kernschätzern ist hier auf die Behandlung von Randproblemen zu achten, da diese insbesondere an den Rändern der Histogrammbins im Innern der Domäne auftreten können.

1.3.5 Bandbreitenbestimmung

Bei vielen nicht-parametrischen Verfahren wie Histogramm- oder Kernschätzern existiert ein sogenannter *Glättungsparameter*, auch *Bandbreite* genannt. Die Güte der Schätzung kann deutlich von der richtigen Wahl dieses Parameters abhängen. In keiner der bisherigen Arbeiten zur Selektivitätsschätzung wurde dieses Problem adressiert. Das Problem der Bestimmung eines Glättungsparameters bei der Selektivitätsschätzung wurde zuerst in [Mannino et al. 88] adressiert. Dort wurde jedoch nur eine einfache Regel (Sturge Regel) vorgeschlagen, die außer der Stichprobengröße keinerlei statistische Kenngrößen der Stichprobe berücksichtigt. In [Blohs-

feld et al. 99] wurden erstmalig Verfahren zur Schätzung einer asymptotisch optimalen Bandbreite angewendet. Diese werden in dieser Arbeit ausführlich diskutiert und verfeinert. Insbesondere werden auch adaptive Verfahren zur Bandbreitenbestimmung betrachtet.

1.3.6 Zusammenfassung

Die meisten bisher bekannten Verfahren zur Selektivitätsschätzung in Datenbanksystemen befassen sich mit Histogrammschätzern als einer speziellen Variante nicht-parametrischer Verfahren. Die verschiedenen Histogrammschätzer unterscheiden sich im Wesentlichen in der Art der Partitionierung der Domäne in Histogrammbins. In keiner der bekannten Arbeiten wird ein expliziter Bezug der Schätzverfahren zur mathematischen Statistik hergestellt. Das Problem der Wahl der Anzahl von Bins wird in kaum einer der Arbeiten betrachtet. Lediglich in [Ioannidis & Poosala 95] finden sich experimentelle Untersuchungen über die optimale Anzahl von Bins bei Häufigkeitshistogrammen auf Daten, die aus einer kleinen Domäne stammen und bzgl. der Häufigkeit sortiert sind. Die dort getroffene Aussage, daß die Wahl der Anzahl der Bins bei sehr schiefen Verteilungen einfach sei und bei nicht-schiefen Verteilungen keine Rolle spiele ist sicher nicht auf die in dieser Arbeit getroffenen Voraussetzungen verallgemeinerbar.

In der vorliegenden Arbeit werden daher erstmalig die nicht-parametrischen Verfahren mit Hilfe der mathematischen Statistik beschrieben und analysiert. Dabei werden insbesondere Histogrammselektivitätsschätzer und Kernselektivitätsschätzer untersucht.

Es lassen sich unterschiedliche Voraussetzungen an die zugrunde liegenden Daten treffen und in der Literatur zur Selektivitätsschätzung in Datenbanksystemen werden - teilweise implizit - unterschiedliche Voraussetzungen angenommen. In dieser Arbeit werden ausschließlich metrische Daten einer großen Domäne betrachtet. Insbesondere wird somit vorausgesetzt, daß auf den Daten eine Ordnung besteht und die zugrunde liegende Dichtefunktion stetig ist. OBdA wird als metrischer Raum der Raum der reellen Zahlen $\Re^d$ mit der euklidischen Metrik als Abstandsmaß gewählt.

Um die Ergebnisse der Experimente vergleichbar zu machen, wurden einige der bisher bekannten Verfahren zur Selektivitätsschätzung in dieser Arbeit reimplementiert und bei den Experimenten als Referenz-Verfahren zum Vergleich verwendet. Im folgenden werden die neu vorgestellten Verfahren und die Referenz-Verfahren für die Experimente aufgelistet.

In den Experimenten referenzierte univariate Verfahren:

- Selektivitätsschätzung aufgrund der Gleichverteilungsannahme
- Direkte Selektivitätsschätzung
- Equi Width- und Equi Depth-Histogrammselektivitätsschätzer
- Max Diff-Histogrammselektivitätsschätzer ([Poosala et al. 96])

Neu vorgestellte univariate Verfahren:

- Average Shifted Histogrammselektivitätsschätzer

- Kernselektivitätsschätzer ([Blohsfeld 98], [Blohsfeld et al. 99])
- Hybridselektivitätsschätzer

In den Experimenten referenzierte multivariate Verfahren:

- Selektivitätsschätzung aufgrund der Gleichverteilungsannahme
- Direkte Selektivitätsschätzung
- Equi Width-Histogrammselektivitätsschätzer
- Equi Depth-Histogrammselektivitätsschätzer nach [Muralikrishna & DeWitt 88]

Neu vorgestellte multivariate Verfahren:

- Average Shifted Histogrammselektivitätsschätzer
- Z-Kurve ([Buskamp 97])
- KD-Baum-Selektivitätsschätzung ([Buskamp 97])
- Voronoi-Selektivitätsschätzung ([Schneider 97])
- Kernselektivitätsschätzer ([Blohsfeld 98])

2. Grundlagen

In diesem Kapitel werden die grundlegenden und für diese Arbeit relevanten Begriffe sowohl aus dem Bereich Datenbanksysteme als auch der mathematischen Statistik eingeführt. Hierfür wesentliche Begriffe sind dabei der Begriff der Selektivitätsschätzung und der Begriff der Dichteschätzung. Zwischen beiden Begriffen wird hier ein Zusammenhang hergestellt, der es erlaubt das Problem der Selektivitätsschätzung mit Hilfe von Methoden der nicht-parametrischen Statistik zu beschreiben. Dabei werden Methoden der nicht-parametrischen Dichteschätzung in diesem Kapitel nur überblicksweise beschrieben und in einem späteren Kapitel ausführlich in Zusammenhang mit der Selektivitätsschätzung dargestellt. Die in dieser Arbeit verwendete Notation wird in einer Tabelle zusammengefaßt. Abschließend werden bisherige Arbeiten zur Selektivitätsschätzung den verschiedenen Methoden zur Selektivitätsschätzung zugeordnet, sowie ihre Vor- und Nachteile dargestellt.

2.1 Selektivitätsschätzung

Wie bereits in Kapitel 1 konstatiert, wird in dieser Arbeit ein relationales Datenbanksystem zugrunde gelegt. Dazu müssen zunächst Begriffe wie Relation, Instanz, Anfrage und Selektivität einer Anfrage geklärt werden.

Eine Relation kann als eine Tabelle angesehen werden, in der Daten bzgl. bestimmter Eigenschaften (eines bestimmten *Attributs*) enthalten sind. In den Spalten der Tabelle sind nur die Werte eines bestimmten Attributs enthalten. Die Werte eines Attributs gehören zu einem bestimmten Wertebereich, auch *Domäne* des Attributs genannt. Es sind nur einfache Attributtypen wie Zahlen oder Zeichenketten zugelassen, nicht aber strukturierte Attributtypen wie z.B. Mengen oder (komplexe) Objekte. Diese Arbeit bezieht sich ausschließlich auf Attribute mit metrischer Domäne, insbesondere auf reellwertige Attribute. Die Zeilen der Tabelle bilden die eigentlichen Objekte oder Beziehungen ab und werden *Tupel* genannt. Formal gesehen läßt sich eine Relation wie folgt definieren.

Definition 2.1 (*Relation*):

> Seien d Attribute $A_1,...,A_d$ gegeben mit $A_i \in D_i$, $i = 1..d$. D_i heißt *Domäne* (*Wertebereich*) von A_i. Dann ist eine *Relation* R definiert als $R \subseteq D_1 \times \ldots \times D_d$. d heißt die *Stelligkeit* der Relation.

Bei Relationen wird genauer unterschieden nach Schema und Instanz einer Relation. Bei einem *Schema* handelt es sich um die Spezifikation einer Relation. Sie sollte in der Regel einmal erstellt und nicht mehr modifiziert werden. Bei der *Instanz* handelt es sich dagegen um eine Ausprägung einer Relation. Insbesondere können zu einer Relation mehrere Instanzen gehören. Die Instanz einer Relation ist i.a., insbesondere bei operationalen Datenbanken mit beliebigen Transaktionen, stetem Wandel unterzogen. In nicht-operationalen Datenbanksystemen wie z.B.

Data-Warehouses erfolgt ein Update der Datenbasis immerhin in der Regel in periodischen Abständen.

Grundsätzlich unterscheidet man drei Typen von Anfragen. Eine *Selektionsanfrage* (engl. *selection query*) ist an eine Bedingung geknüpft, die sich auf ein Attribut der Relation bezieht. Gesucht werden alle Tupel einer Relation (bzw. deren Instanz), die die Bedingung erfüllen. Der häufigste Fall von Selektionsanfragen sind die *Bereichsanfragen* (engl. *range query*), bei den die Bedingung sich auf einen angegebenen Bereich der Domäne des Attributs bezieht. Hiervon sind wiederum *Fensteranfragen*, bei denen der Anfragebereich aus einem (gegebenenfalls mehrdimensionalen) Intervall besteht, der häufigste Fall. Eine andere Bereichsanfrage wäre z.B. die *k-nächste Nachbarn Anfrage* (engl. *k-nearest neighbor query*), bei der die k Tupel gesucht werden, die von allen Tupeln der Relation den geringsten Abstand bzgl. einer gegebenen Metrik zu einem Anfragepunkt haben. Ein weiterer Anfragetyp vom Typ der Selektionsanfrage wäre z.B. eine *Punktanfrage*, bei der alle diejenigen Tupel gesucht werden, die in einem angegebenen Attribut mit dem Anfragepunkt übereinstimmen. Neben den Selektionsanfragen gibt es noch Projektionsanfragen und Join-Anfragen. Bei einer *Projektionsanfrage* werden nur bestimmte Attribute einer Relation zurückgegeben. Eine *Join-Anfrage* dagegen verknüpft zwei Relationen über ein gemeinsames Schlüssel-Attribut, so daß es sich hierbei um einen zweistelligen Operator handelt. Auch hier werden mehrere Untertypen unterschieden. Beim *Equi-Join* werden z.B. alle die Tupel miteinander verbunden, die der Gleichheitsbedingung bzgl. eines gemeinsamen Attributs genügen.

Diese Arbeit behandelt die Selektivitätsschätzung von Bereichsanfragen als eine der wichtigsten und häufigsten dieser verschiedenen Anfragetypen.

Definition 2.2 (*Bereichsanfrage*):

> Sei eine Relation R mit Attributen aus $\Re^d$ gegeben. Eine *Bereichsanfrage* $Q = Q(a,b)$ mit $a, b \in \Re^d$, $a<b$, auf R sei hier allgemein definiert als ein abgeschlossener, achsenparalleler d-dimensionaler Hyperwürfel $[a,b]$. Falls aus dem Zusammenhang klar ist, daß es sich um eine Bereichsanfrage handelt, wird im folgenden auch einfach nur der Begriff *Anfrage* verwendet.

Beispiel: Eine eindimensionale Bereichsanfrage wäre z.B. die Frage danach, wieviele Arbeitnehmer in der Bundesrepublik zwischen 50.000,- und 100.000,- DM brutto im Jahr verdienen. Eine zweidimensionale Bereichsanfrage wäre z.B. die Frage, wieviele Städte sich zwischen (oder auf) dem a_1-ten und b_1-ten Breitengrad und zwischen dem a_2-ten und b_2-ten Längengrad befinden.

Dieser Typ Bereichsanfrage heißt häufig auch Fensteranfrage. Prinzipiell ließen sich auch andere Bereichsanfragen definieren. Eine öfters auftretende Bereichsanfrage wäre z.B. die Frage danach, wieviele Objekte sich innerhalb einer bestimmten Entfernung r eines anderen Objektes befinden. Der Bereich wäre demnach z.B. (bei entsprechender Metrik) eine Kugel um ein gegebenes Objekt mit vorgegebenem Radius r. Prinzipiell wäre eine solche Anfrage mit den

gleichen Methoden lösbar wie bei den hier definierten Fensteranfragen vorgeschlagen. Die direkte Berechnung der Schnittmengen bzw. des Integrales zur Selektivitätsschätzung gestaltet sich jedoch technisch ungleich schwieriger. Diese Arbeit beschränkt sich daher auf die am häufigsten verwendete Art von Bereichsanfragen, die Fensteranfragen.

Für Grundbegriffe der mathematischen Statistik und Wahrscheinlichkeitsrechnung wie z.B. Zufallsvariable und Wahrscheinlichkeitsmaß sei auf die einschlägige Literatur verwiesen ([Bauer 91], [Bosch 89], [Hartung & Elpert 95], [Rasch 95] u.a.). Einige wichtige Definitionen seien im folgenden noch einmal aufgeführt, da sie sich durch die gesamte Arbeit ziehen. Grundlegend für die Selektivitätsschätzung sind die Begriffe Dichtefunktion und Verteilungsfunktion.

Sei $\tilde{\Re} = \Re \cup \{\pm\infty\}$. Im folgenden wird unter einer *Zufallsvariable X*: $\Omega \to \tilde{\Re}^d$ immer eine numerische Zufallsvariable bzgl. des Wahrscheinlichkeitsraumes (Ω,*A*,*P*) mit Wahrscheinlichkeitsmaß *P* verstanden.

Definition 2.3 (*Dichtefunktion, Verteilungsfunktion*):

Sei *X:* $\Omega \to \tilde{\Re}^d$ eine numerische Zufallsvariable bzgl. des Wahrscheinlichkeitsraumes (Ω,*A*,*P*) mit Wahrscheinlichkeitsmaß *P*. Dann ist die *Verteilungsfunktion* $F_X = F_X(x)$ von *X* an der Stelle $x = (x_1, \ldots, x_d)^t \in \tilde{\Re}^d$ definiert als die Wahrscheinlichkeit, daß *X* kleiner oder gleich *x* ist:

$$F_X : \tilde{\Re}^d \to [0,1],\ F_X(x) = P(X_1 \le x_1, \ldots, X_d \le x_d). \tag{2.1}$$

Falls klar ist, welche Zufallsvariable gemeint ist, wird statt F_X auch einfach *F* verwendet.

Die Zufallsvariable *X* ist stetig verteilt, falls eine reelle nicht-negative integrierbare Funktion *f* existiert, so daß für alle $x = (x_1, \ldots, x_d)^t \in \tilde{\Re}^d$ gilt:

$$F(x) = \int_{-\infty}^{x_d} \ldots \int_{-\infty}^{x_1} f(t)dt_1 \ldots dt_d. \tag{2.2}$$

f heißt *Dichtefunktion.*

Bemerkung 2.1:

Für die Dichtefunktion gilt insbesondere $f(x) \ge 0 \forall x \in \Re^d$ und $\int_{-\infty}^{\infty} f(x)dx = 1$.

Die Dichtefunktion spielt in dieser Arbeit eine zentrale Rolle bei der Definition der Selektivität.

Im folgenden wird bei den mathematischen Betrachtungen davon ausgegangen, daß die Dichtefunktion stetig ist. In Kapitel 3.5 und in den Experimenten in Kapitel 5 werden abweichend davon Dichtefunktionen mit Sprungstellen betrachtet.

Als Beispiel zeigt Abbildung 2.1 die Dichtefunktion (durchgezogene Kurve) und die Verteilungsfunktion (gepunktete Kurve) der Standardnormalverteilung $N(0,1)$ mit $f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$.

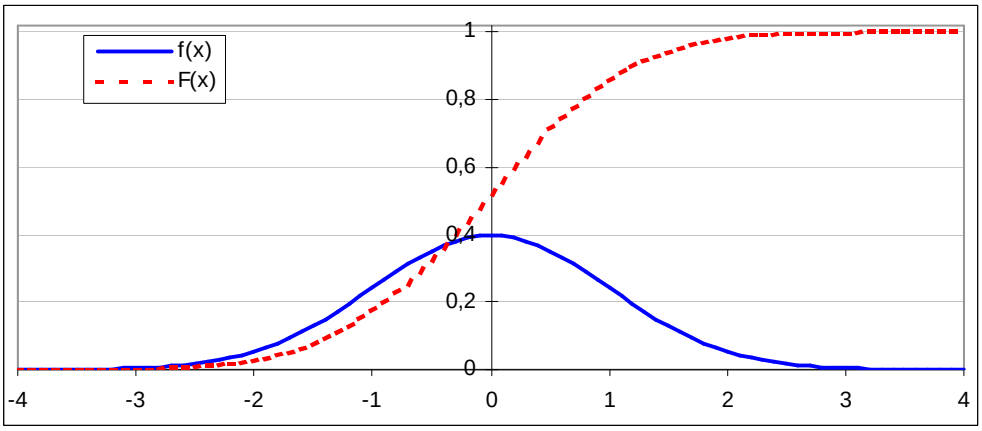


Abbildung 2.1: Dichtefunktion *f(x)* und Verteilungsfunktion *F(x)* der Standardnormalverteilung

Da es i.a. in der Praxis nicht praktikabel ist, die gesamte Datenbankinstanz (Grundgesamtheit) zur Selektivitätsschätzung heranzuziehen, ist das Ziehen einer repräsentativen Stichprobe aus der Grundgesamtheit erforderlich. Ein weiterer in dieser Arbeit wichtiger statistischer Begriff ist daher der Begriff der Zufallsstichprobe. Eine Zufallsstichprobe $\{X_1,..., X_n\}$ heißt *einfach*, wenn die Zufallsvariablen $X_1,..., X_n$ *unabhängig* sind und alle dieselbe Verteilungsfunktion F besitzen (*identisch verteilt* sind). Im folgenden ist mit Zufallsstichprobe immer eine einfache Zufallsstichprobe gemeint.

In dieser Arbeit werden im folgenden die Begriffe Schätzer und Schätzfunktion synonym gebraucht. Es werden ausschließlich Dichteschätzer bzw. die daraus abgeleiteten Selektivitätsschätzer betrachtet (s.u.).

Der zentrale Begriff der Selektivität wird in dieser Arbeit unterschieden zum einen nach der Instanz-Selektivität, d.h. der aktuellen Selektivität einer konkreten Datenbank-Instanz, und zum anderen nach der Relation-Selektivität, d.h. der Selektivität der einer Instanz zugrundeliegenden Datenbank-Relation. Letztere gibt die Wahrscheinlichkeit wieder, daß sich ein Element der Relation in dem Anfragebereich befindet. Dagegen bedeutet die Instanz-Selektivität die Anzahl der Elemente eines Anfragebereichs der aktuellen Instanz.

Definition 2.4 (*Instanz-Selektivität*):

Seien eine Relation R und eine Instanz $I = I_R(N)$ mit N Tupeln gegeben. Die *Instanz-Selektivität* $\sigma_I(Q)$ einer Bereichsanfrage $Q = Q(a,b)$ ist definiert als die Menge der Tupel der Instanz, die im Anfragebereich liegen:

$$\sigma_I(a, b) = |\{x \in I: a \le x \le b\}| \tag{2.3}$$

Für $d > 1$ gelten die Beziehungen komponentenweise, d.h.

$$\sigma_I(a, b) = \left|\{x \in I: a_i \le x_i \le b_i \forall i = 1 \ldots d\}\right|$$

Definition 2.5 (*Relation-Selektivität*):

Sei eine Relation R gegeben. Sei weiter X eine unabhängige Zufallsvariable. Die *Relation-Selektivität* $\sigma(Q) = \sigma(a, b)$ einer Bereichsanfrage $Q = Q(a,b)$ ist definiert als die Wahrscheinlichkeit, daß $X \in R$ in dem Anfragebereich liegt:

$$\sigma_R(a, b) = P(a \le X \le b). \tag{2.4}$$

Für $d > 1$ gelten die Beziehungen wieder komponentenweise.

Wenn nichts weiteres gesagt ist, so ist mit *Selektivität* σ in dieser Arbeit die Relation-Selektivität σ_R gemeint.

Die Instanz-Selekivität ist dabei die in Datenbanksystemen interessierende Größe, während die Relation-Selektivität als Modell für die Instanz-Selektivität gewählt wird. Die Instanz-Selektivität ist hierbei (im Unterschied zu z.B. [Kemper & Eickler 97]) als absolute Größe definiert, während die Relation-Selektivität als relative Größe definiert ist. In der bisherigen Literatur im Bereich der Selektivitätsschätzung von Datenbankanfragen wird Selektivität immer als Instanz-Selektivität definiert, so daß die hier getroffene Unterscheidung neu ist. Einige Autoren vernachlässigen es weiterhin auf den Unterschied von relativer und absoluter (Instanz-) Selektivität hinzuweisen. Dabei verwenden verschiedene Autoren verschiedene Definitionen von Selektivität. [Whang et al. 94] definieren Selektivität explizit als relative (Instanz-) Selektivität. [Poosala & Ioannidis 97] berechnen die Ergebnisgröße (result size) und damit die absolute (Instanz-) Selektivität. [Muralikrishna & DeWitt 88] bestimmen die Anzahl der Tupel in der Anfrageregion, was ebenfalls der absoluten (Instanz-) Selektivität entspricht. Dabei läßt sich die absolute Instanz-Selektivität jederzeit in die relative Instanz-Selektivität durch Division durch die Größe der Grundgesamtheit N überführen. Die folgenden Eigenschaften der Relation- und der Instanz-Selektivität rechtfertigen die Wahl der Relation-Selektivität als Modell für die Instanz-Selektivität unter Berücksichtigung der Unterscheidung von absoluter und relativer Selektivität.

Die Relation-Selektivität σ_R sollte intuitiv die folgenden Eigenschaften besitzen. Diese folgen direkt aus der Definition der Wahrscheinlichkeit.

- Die Relation-Selektivität liegt zwischen Null und Eins: $0 \le \sigma_R(Q) \le 1 \quad \forall Q$.
- Die Relation-Selektivität einer leeren Anfrage ist Null: $\sigma_R(\varnothing) = 0$.

- Deckt die Anfrage die gesamte Domäne ab, so ist die Relation-Selektivität gleich Eins: $\sigma_R(X) = 1$.
- Ist eine Anfrage Q_1 gegeben, die eine echte Teilmenge einer zweiten Anfrage Q_2 ist, dann ist die Relation-Selektivität der Anfrage Q_1 kleiner oder gleich der Relation-Selektivität von Q_2: $Q_1 \subset Q_2 \Rightarrow \sigma_R(Q_1) \leq \sigma_R(Q_2)$. (Monotonieeigenschaft)

Die Instanz-Selektivität σ_I besitzt die folgenden Eigenschaften, die direkt aus der Definition folgen.

- Die Instanz-Selektivität liegt zwischen Null und der Anzahl *N* der Elemente der Instanz: $0 \leq \sigma_I(Q) \leq N \quad \forall Q$
- Die Instanz-Selektivität einer leeren Anfrage ist Null: $\sigma_I(\varnothing) = 0$.
- Deckt die Anfrage die gesamte Domäne ab, so ist die Instanz-Selektivität gleich der Anzahl *N* der Elemente der Instanz: $\sigma_I(X) = N$.
- Ist eine Anfrage Q_1 gegeben, die eine echte Teilmenge einer zweiten Anfrage Q_2 ist, dann ist die Instanz-Selektivität der Anfrage Q_1 kleiner oder gleich der Instanz-Selektivität von Q_2: $Q_1 \subset Q_2 \Rightarrow \sigma_I(Q_1) \leq \sigma_I(Q_2)$. (Monotonieeigenschaft)

Durch die Definition der Selektivität mit Hilfe der Wahrscheinlichkeitsfunktion ist ein direkter Zusammenhang mit der mathematischen Statistik gegeben. Die Relation-Selektivität kann nun elegant mit Hilfe der Definition der Dichtefunktion ausgedrückt werden:

$$\sigma(a, b) = \int_a^b f(t)dt \tag{2.5}$$

Im Falle der Dimension *d* = 1 gilt unter Verwendung der Verteilungsfunktion weiterhin:

$$\sigma(a, b) = \int_a^b f(t)dt = F(b) - F(a) \tag{2.6}$$

Dies läßt sich in dieser einfachen Form nicht für höhere Dimension *d* > 1 ableiten. Unter der Voraussetzung, daß es sich bei *X* um eine unabhängige Zufallsvariable handelt, gilt vielmehr (vgl. [Bauer 91]):

$$\sigma(a, b) = \int_a^b f(t)dt = \prod_{i=1}^{d} (F_{X_i}(b_i) - F_{X_i}(a_i)) \tag{2.7}$$

Man nennt das Produkt auf der rechten Seite auch *d-dimensionale Differenz.*

Für *d* = 2 ergibt sich z.B.:

$$\sigma(a,b) = \int_a^b f(t)dt = \int_{a_1}^{b_1}\int_{a_2}^{b_2} f(t_1,t_2)dt_2dt_1$$

$$= \int_{a_1}^{b_1}\left(\int_{-\infty}^{b_2} f(t_1,t_2)dt_2 - \int_{-\infty}^{a_2} f(t_1,t_2)dt_2\right)dt_1$$

$$= \int_{a_1}^{b_1}\int_{-\infty}^{b_2} f(t_1,t_2)dt_2dt_1 - \int_{a_1}^{b_1}\int_{-\infty}^{a_2} f(t_1,t_2)dt_2dt_1$$

$$= \int_{-\infty}^{b_1}\int_{-\infty}^{b_2} f(t_1,t_2)dt_2dt_1 - \int_{-\infty}^{b_1}\int_{-\infty}^{a_2} f(t_1,t_2)dt_2dt_1 - \int_{-\infty}^{a_1}\int_{-\infty}^{b_2} f(t_1,t_2)dt_2dt_1 + \int_{-\infty}^{a_1}\int_{-\infty}^{a_2} f(t_1,t_2)dt_2dt_1$$

$$= F(b_1,b_2) - F(b_1,a_2) - F(a_1,b_2) + F(a_1,a_2)$$

$$= F_{X_1}(b_1)F_{X_2}(b_2) - F_{X_1}(b_1)F_{X_2}(a_2) - F_{X_1}(a_1)F_{X_2}(b_2) + F_{X_1}(a_1)F_{X_2}(a_2)$$

$$= (F_{X_1}(b_1) - F_{X_1}(a_1))(F_{X_2}(b_2) - F_{X_2}(a_2)).$$

Im folgenden sei ein einfaches Beispiel für den Zusammenhang zwischen Selektivität und Dichtefunktion im univariaten Fall gegeben. Dazu sei eine Relation mit 100.000 eindimensionalen Tupeln gegeben. Die Werte seien gleichverteilt zwischen 0 und 100. Dann ist die Dichtefunktion eine horizontale Linie mit konstantem Wert 1/100 zwischen 0 und 100 ($f(x) = \begin{cases} 1/100 & \text{falls } 0 \le x \le 100 \\ 0 & \text{sonst} \end{cases}$). Sei nun eine Anfrage $Q(20,30)$ gegeben. Dann läßt sich das Integral $\int_{20}^{30} f(t)dt$ trivial berechnen als $(30-20)\cdot 1/100$, was eine Selektivität von 0,1 ergibt. Dies bedeutet, daß 10% der Daten im Intervall [20,30] liegt, wie zu erwarten war, vgl. Abbildung 2.2.

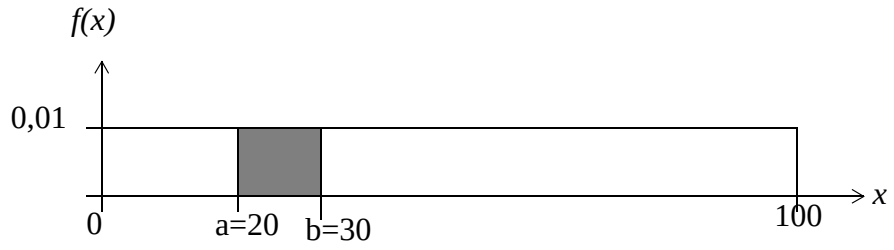


Abbildung 2.2: Beispiel zur Selektivität bei gleichverteilten Daten

Die Instanz einer Relation kann als sehr große Stichprobe aus der Relation selbst angesehen werden. Je nachdem, ob Duplikate in der Instanz zugelassen sind oder nicht, kann die Instanz als Stichprobe gezogen mit Zurücklegen (Duplikate zugelassen) oder ohne Zurücklegen (Duplikate nicht zugelassen) angesehen werden. In dieser Arbeit wird die Instanz einer Relation als Stichprobe gezogen mit Zurücklegen betrachtet, d.h. also daß Duplikate in der Instanz zugelassen sind. Da es sich bei der Instanz um eine Stichprobe aus der Relation handelt, ergibt sich als Beziehung zwischen Relation- und Instanz-Selektivität ein Stichprobenfehler. Dieser kann bei genügend großer Instanz als relativ gering angesehen werden.

Im allgemeinen ist die Anzahl N der Tupel in einer Instanz zu groß um Schätzmethoden direkt auf der Instanz durchzuführen. Stattdessen wird aus der Instanz eine weitere kleinere Zufallsstichprobe durch Ziehen ohne Zurücklegen gezogen. Auf die verschiedenen Verfahren zum Ziehen eine Stichprobe sei hier nur kurz hingewiesen. Von Vitter stammen einige interessante Verfahren für den Datenbank-Bereich. Dazu gehört eine Variante des Reservoir Sampling, welches dann sinnvoll ist, wenn die Größe der Datenbank nicht im vorhinein bekannt ist [Vitter 85]. In den Experimenten wurden i.a. einfache Stichproben ohne Zurücklegen der Größe $n = 2.000$ gezogen (siehe Kapitel 5), was z.B. mit den Experimenten von Poosala [Poosala et al. 96] übereinstimmt. Vergleichsweise wurden auch andere Stichprobengrößen untersucht. Im weiteren ist mit Stichprobe immer die Stichprobe gemeint, die aus der Instanz gezogen wurde und nicht die Instanz selbst.

Aufgrund der Gleichungen (2.5) und (2.6) für die Selektivität ergeben sich nun zwei Möglichkeiten der Selektivitätsschätzung $\hat{\sigma}$, indem man die Funktion der wahren Dichte f durch eine Dichteschätzung $\hat{f} = \hat{f}(X;x)$ mit Zufallsstichprobe $X = (X_1, \ldots, X_n)$ ersetzt:

1. Schätzung der Selektivität anhand einer Schätzung der Dichtefunktion und anschließender Integration:

$$\hat{\sigma}(a, b) = \int_a^b \hat{f}(t)dt \tag{2.8}$$

2. Schätzung der Selektivität anhand einer Schätzung der Verteilungsfunktion. Dies wäre für $d = 1$:

$$\hat{\sigma}(a, b) = \hat{F}(b) - \hat{F}(a) \tag{2.9}$$

2.2 Gütekriterien und Fehlermaße

In diesem Abschnitt werden in der Statistik übliche Fehlermaße und Gütekriterien zur Bewertung von Schätzern vorgestellt. Dabei kommen sowohl punktweise als auch die gesamte Domäne betrachtende Fehlermaße zum Tragen. Besonderer Wert wird auf Aussagen bei sehr großem Stichprobenumfang gelegt. Asymptotische Betrachtungen erlauben es weiterhin die

verschiedenen nicht-parametrischen Schätzer bzgl. ihres Konvergenzverhaltens miteinander zu vergleichen.

Um die Ergebnisse der Experimente zur Selektivitätsschätzung mit verschiedenen künstlichen und realen Testdaten miteinander sowie mit Ergebnissen aus der Datenbank-Literatur zur Selektivitätsschätzung vergleichbar zu machen (siehe Kapitel 5), werden des weiteren die hierfür üblichen Fehlermaße des mittleren absoluten und relativen Fehlers eines Selektivitätsschätzers herangezogen.

Definition 2.6 (*erwartungstreu*, *Bias, Varianz*):

Sei f eine Funktion. Eine Schätzfunktion $\hat{f}$ für f heißt *erwartungstreu* (*unverzerrt, unbiased*), wenn der Erwartungswert der Schätzfunktion $\hat{f}$ existiert und gleich der zu schätzenden Funktion f ist:

$$E(\hat{f}(x)) = f(x) \tag{2.10}$$

Der *Bias* (die *Verzerrung*) des Schätzers ist dann definiert als

$$Bias(\hat{f}(x)) = E(\hat{f}(x)) - f(x) \tag{2.11}$$

Mit anderen Worten, der Schätzer ist erwartungstreu, wenn sein Bias gleich 0 ist.

Die *Varianz Var* eines Schätzers ist definiert als

$$Var(\hat{f}(x)) = E(\hat{f}^2(x)) - E(\hat{f}(x))^2 \tag{2.12}$$

Ein in der Statistik übliches Maß für die Güte eines Schätzers in einem Punkt ist der mittlere quadratische Fehler:

Definition 2.7 (*mittlerer quadratischer Fehler, MSE*):

Seien eine Funktion f und ihre Schätzfunktion $\hat{f}$ gegeben. Dann ist der *mittlere quadratische Fehler* (*mean squared error, MSE*) definiert als

$$MSE(\hat{f}(x)) = E((\hat{f}(x) - f(x))^2) \tag{2.13}$$

Bemerkung 2.2:

Der MSE läßt sich auch eleganter schreiben als

$$MSE(\hat{f}(x)) = E((\hat{f}(x) - f(x))^2) = Var(\hat{f}(x)) + Bias(\hat{f}(x))^2 \tag{2.14}$$

mit Bias und Varianz wie in Gleichungen (2.11) und (2.12) definiert.

Diese Aufsplittung in Varianz und Bias wird im Zusammenhang mit der Bandbreitenbestimmung in Kapitel 4 verwendet, um den unterschiedlichen Einfluß der beiden Größen zu beleuchten.

Ist der Schätzer $\hat{f}$ erwartungstreu, so gilt: $MSE(\hat{f}) = Var(\hat{f})$

Bei dem mittleren quadratischen Fehler handelt es sich um ein lokales Maß, das die Abweichung der Schätzfunktion von der Zielfunktion in einem Punkt mißt. Um den globalen Fehler über der gesamten Domäne zu messen, benötigt man ein Maß für die Abweichung der Schätzfunktion $\hat{f}(.)$ von der Funktion $f(.)$, die zudem nicht auf eine spezielle Realisation einer Stichprobe festgelegt ist. Ein solches Maß ist durch den mittleren integrierten quadratischen Fehler gegeben, der wie folgt definiert ist:

Definition 2.8 (*mittlerer integrierter quadratischer Fehler, MISE*):

Sei eine Funktion f mit Schätzfunktion $\hat{f}$ gegeben. Dann ist der *mittlere integrierte quadratische Fehler* (*mean integrated squared error,* MISE) definiert als

$$MISE(\hat{f}) = E\left(\int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 dx \right). \tag{2.15}$$

In Anlehnung an die Literatur wird im folgenden der abkürzende Begriff *MISE* verwendet.

Bemerkung 2.3:

Der MISE läßt sich auch definieren als M(ISE), wobei der ISE hier als die L_2-Norm definiert ist. Es ließe sich auch jede andere beliebige L_p-Norm ($p = 1,...,\infty$) verwenden, aber die L_2-Norm zeichnet sich durch eine einfachere mathematische Behandlung aus. Für weitere Alternativen siehe z.B. [Scott 92].

Aufgrund der Eigenschaft der Nicht-Negativität der Schätzfunktion und der Dichtefunktion f läßt sich der MISE mit Hilfe von Varianz und Bias eleganter schreiben als (vgl. [Silverman 86], S.35f.):

$$MISE(\hat{f}) = \int_{-\infty}^{\infty} Var(\hat{f}(X))dX + \int_{-\infty}^{\infty} Bias(\hat{f}(X))^2 dX \tag{2.16}$$

Ein in dieser Arbeit weiterer wichtiger Begriff ist der Begriff der Konvergenzordnung. Er gestattet es, Aussagen über die Geschwindigkeit der Konvergenz (eines Schätzers) zu machen.

Definition 2.9 (*Konvergenzordnung*):

Seien zwei reellwertige Folgen a_n und b_n gegeben.

a_n heißt von der *Konvergenzordnung* $O(b_n)$ *(für* $n \to \infty$*)*, falls $\limsup_{n \to \infty} (a_n / b_n) < \infty$.

a_n heißt von der *Konvergenzordnung* $o(b_n)$ *(für* $n \to \infty$*)*, falls $\lim_{n \to \infty} (a_n / b_n) = 0$.

a_n und b_n heißen *asymptotisch (äquivalent),* falls $\lim_{n \to \infty} (a_n / b_n) = 1$.

Definition 2.10 (*konsistenter Schätzer*):

Sei $\hat{f} = \hat{f}_n$ eine Folge von Schätzfunktionen für f und $P(A)$ die Wahrscheinlichkeit für ein Ereignis A. $\hat{f}$ heißt (*schwach*) *konsistent,* falls für alle $\varepsilon > 0$ gilt:

$$\lim_{n \to \infty} P(|f - \hat{f}_n| > \varepsilon) = 0 \tag{2.17}$$

$\hat{f}$ heißt *konsistent im mittleren Fehlerquadrat* (oder: *stark konsistent*), wenn gilt

$$\lim_{n \to \infty} E((f - \hat{f}_n)^2) = 0 \tag{2.18}$$

Konsistenz eines Schätzers besagt mit anderen Worten, daß der Schätzer sich mit hoher Wahrscheinlichkeit der zu schätzenden Funktion mit beliebiger Genauigkeit annähert, wenn nur die Anzahl der Stichprobenelemente genügend groß wird. Somit ist Konsistenz eine entscheidende Forderung an einen Schätzer, die auch in dieser Arbeit berücksichtigt wird. Aus der starken Konsistenz folgt direkt die schwache Konsistenz.

Bemerkung 2.4:

Ein Kriterium für Konsistenz läßt sich aus Bias und Varianz des Schätzers ableiten; die Behauptung folgt direkt aus obiger Definition:

Sei $\hat{f} = f_n$ ein erwartungstreuer Schätzer für die Funktion f und es gelte

$$\lim_{n \to \infty} Var(f_n) = 0. \tag{2.19}$$

Dann ist f_n konsistent für f.

In vielen praktischen Anwendungen kann der MISE nicht exakt berechnet werden. In diesen Fällen kann man sich auf den *asymptotischen MISE* (*AMISE*) zurückziehen. Hierzu wird der MISE mittels der Taylorentwicklung [Forster 79] von f in eine Reihe entwickelt und die Terme werden ab einer gewissen Ordnung abgeschnitten. Der AMISE findet in dieser Arbeit Verwendung bei der Berechnung der asymptotisch optimalen Bandbreite und der Berechnung der Konvergenzrate in Kapitel 4.

In Kapitel 5 werden die theoretischen Ergebnisse mit Hilfe von Experimenten zur Schätzung der Selektivität bei einer gegebenen Instanz I mit N Tupeln validiert. Hierbei wird die Schätzung der Relation-Selektivität mit der Instanz-Selektivität verglichen. Somit lassen sich absoluter und relativer Fehler auf die Selektivitätsschätzung bei vorgegebener Instanz einer Relation definieren. Im folgenden sei zunächst der Fehler für eine einzelne Anfrage definiert und anschlie-

ßend der mittlere Fehler für eine ganze Menge von Anfragen. Dabei wird der Fehler zwischen wahrer Instanz-Selektivität und geschätzter Relation-Selektivität der Einfachheit halber Selektivitätsfehler genannt.

Definition 2.11 (*absoluter/relativer Selektivitätsfehler*):

Sei eine Relation R gegeben sowie eine Instanz I dieser Relation mit N Tupeln. Sei weiterhin eine Anfrage Q gegeben mit der Instanz-Selektivität $\sigma_I(Q)$, sowie ein Selektivitätsschätzer $\hat{\sigma}_R$.

Dann ist der *absolute Selektivitätsfehler* $h(Q) = \eta_{R,I}(\sigma_I(Q), \hat{\sigma}_R(Q))$ definiert als die absolute Differenz von der Instanz-Selektivität $\sigma_I(Q)$ der Anfrage Q und der geschätzten Relation-Selektivität $\hat{\sigma}_R(Q)$ von Q mal der Anzahl der Tupel N:

$$\eta_{R,I}(\sigma_I(Q), \hat{\sigma}_R(Q)) := \left|\sigma_I(Q) - \hat{\sigma}_R(Q)N\right| \tag{2.20}$$

Damit der absolute Fehler vergleichbar bzgl. verschiedener Instanzen (mit gegebenenfalls unterschiedlich vielen Tupeln N) ist, läßt sich der *absolute Selektivitätsfehler relativ zur Instanzgröße* definieren als

$$\eta_{R,I(N)}(\sigma_I(Q), \hat{\sigma}_R(Q)) := \left|\sigma_I(Q)/N - \hat{\sigma}_R(Q)\right| \tag{2.21}$$

Der *relative Selektivitätsfehler* ist nun analog - für $\sigma_I(Q) > 0$ - definiert als

$$\varepsilon_{R,I}(\sigma_I(Q), \hat{\sigma}_R(Q)) := \frac{\left|\sigma_I(Q) - \hat{\sigma}_R(Q)N\right|}{\sigma_I(Q)} \tag{2.22}$$

Wenn klar ist, auf welche Relation oder Instanz oder Anfrage sich der Fehler bezieht, wird der entsprechende Ausdruck im folgenden auch weggelassen.

Man beachte, daß der relative Fehler für $\sigma_I(Q) = 0$ nicht definiert ist. Dies spielt für die Experimente in Kapitel 5 jedoch keine Rolle, da die Mittelpunkte der Anfragebereiche als Stichprobeelemente aus der gegebenen Datenbankinstanz gewählt wurden. Somit liegt also immer mindestens ein Element im Anfragebereich, nämlich der Mittelpunkt der Anfrage selbst. Daraus folgt, daß in den Experimenten $\sigma_I(Q_i) > 0$ für alle $i = 1..m$ ist.

Definition 2.12 (*mittlerer relativer/absoluter Selektivitätsfehler*):

Seien R eine Relation mit der Instanz I der Größe N sowie Q_i, $i = 1..m$, $\sigma_I(Q_i) > 0$, eine endliche Anzahl von Anfragen. Dann sind der *mittlere absolute* (*relativ zur Instanzgröße*) bzw. *mittlere relative Selektivitätsfehler*, $\bar{\eta} = \bar{\eta}(Q_{i,\,i=1\ldots m}) = (\overline{\eta_{R,I}}(\sigma_I, \hat{\sigma}_R, Q_{i,\,i=1\ldots m})$, $\overline{\eta_N} = \overline{\eta_N}(Q_{i,\,i=1\ldots m}) = \overline{\eta_{R,I(N)}}(\sigma_I, \hat{\sigma}_R, Q_{i,\,i=1\ldots m})$, $\bar{\varepsilon} = \bar{\varepsilon}(Q_{i,\,i=1\ldots m}) = \overline{\varepsilon_{R,I}}(\sigma_I, \hat{\sigma}_R, Q_{i,\,i=1\ldots m})$, jeweils definiert als das Mittel der

absoluten (relativ zur Instanzgröße) bzw. relativen Einzel-Selektivitätsfehler aller Anfragen Q_i:

$$\bar{\eta}(Q_{i,\, i=1\ldots m}) = \frac{1}{m} \cdot \sum_{i=1}^{m} \eta(Q_i), \tag{2.23}$$

$$\overline{\eta_N}(Q_{i,\, i=1\ldots m}) = \frac{1}{m} \cdot \sum_{i=1}^{m} \eta_N(Q_i), \tag{2.24}$$

$$\bar{\varepsilon}(Q_{i,\, i=1\ldots m}) = \frac{1}{m} \cdot \sum_{i=1}^{m} \varepsilon(Q_i). \tag{2.25}$$

In dieser Arbeit wird i.a. der relevantere relative Fehler zum Vergleich der verschiedenen Methoden zur Selektivitätsschätzung herangezogen. In den meisten Fällen entsprechen sich die Ergebnisse bzgl. absolutem und relativem Fehler bei den verschiedenen Methoden. Allerdings kann es z.B. bei besonders schiefen Verteilungen wie bei der Exponentialfunktion zu Verzerrungen kommen, vgl. hierzu Kapitel 5.

2.3 Nichtparametrische Schätzverfahren

2.3.1 Empirische Verteilungsfunktion

Zur einfachen Schätzung der Verteilungsfunktion F kann die empirische Verteilungsfunktion F_n benutzt werden. Deren Wert an einer Stelle x ist bei gegebener Stichprobengröße n definiert als die Anzahl derjenigen Stichprobenelemente, die kleiner oder gleich der auszuwertenden Stelle x sind:

Definition 2.13 (*Empirische Verteilungsfunktion*):

Sei $\{X_1, \ldots, X_n\}$ eine einfache Zufallsstichprobe der Verteilung F. Dann ist die *empirische Verteilungsfunktion* F_n wie folgt definiert:

$$F_n(x) = \frac{\left|\{X_i \le x\}\right|}{n},\ i=1,\ldots,n \tag{2.26}$$

Im Falle der Dimension $d > 1$ sind die x und X_i Vektoren aus $\Re^d$ und die Ungleichung $X_i \le x$ ist komponentenweise zu verstehen.

Im Falle $d = 1$ gilt:

$$F_n(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} 1_{(-\infty,\, x]}(X_i) \text{ mit } 1_C(x) = \begin{cases} 1 \text{ falls } x \in C \\ 0 \text{ falls } x \notin C \end{cases}. \tag{2.27}$$

Ein Beispiel für eine univariate Schätzung der Verteilung mittels empirischer Verteilungsfunktion findet sich in Abbildung 2.4. Geschätzt wurde die Standardnormalverteilung im Bereich $[-4, 4]$ auf einer Stichprobe der Größe $n = 50$.

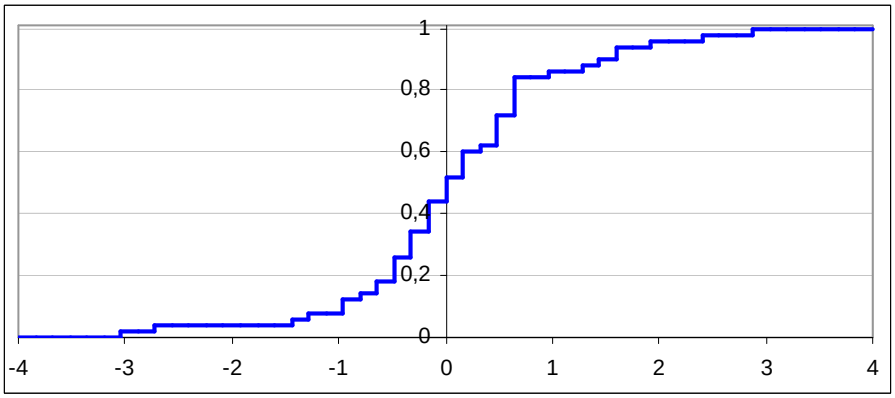


Abbildung 2.3: Beispiel für Schätzung der wahren Verteilung mit empirischer Verteilungsfunktion.

Es kann mit dem Satz von Glivenko-Cantelli [Bauer 91] gezeigt werden, daß die empirische Verteilungsfunktion gleichmäßig gegen die wahre Verteilungsfunktion konvergiert: $P(\sup(|F_n(x) - F(x)|) \to 0) = 1$. Zudem gilt für den Erwartungswert $E(F_n(x)) = E(1_{(-\infty, x]}(X)) = \int_{-\infty}^{\infty} 1_{(-\infty, x]}(X) f(X) dX = \int_{-\infty}^{x} f(X) dX = F(x)$, d.h. die empirische Verteilungsfunktion ist für $d = 1$ erwartungstreu. Ebenso kann für d>1 gezeigt werden, daß die empirische Verteilungsfunktion erwartungstreu ist. Weiterhin gilt für die Varianz: $Var(F_n(x)) = (F(x)(1 - F(x)))/n$. Mit Bemerkung 2.4 folgt somit sofort, daß es sich bei der empirischen Verteilungsfunktion um einen konsistenten Schätzer handelt.

Die Schätzung mittels der empirischen Verteilungsfunktion zeichnet sich vor allem durch ihre Einfachheit aus. Leider ist die empirische Verteilungsfunktion eine Treppenfunktion und damit i.a. keine analytisch stetige Funktion. Die zugrunde liegende wahre Verteilungsfunktion kann jedoch durchaus (analytisch) stetig sein. Ein weiterer Nachteil im Vergleich zu anderen nichtparametrischen Schätzern ist, daß aufgrund der langsamen Konvergenzordnung ($O(n^{-1/2})$) i.a. zu große Stichprobenmengen benötigt werden.

2.3.2 Histogrammdichteschätzer

Grundlage eines jeden Histogrammes ist die vollständige Partitionierung des Datenraums in $k < \infty$ disjunkte Bereiche C_i, $i = 1..k$, üblicherweise *Bins* genannt, wobei den einzelnen Bins charakteristische Eigenschaften zugeordnet werden. Bei Histogrammen zur Dichteschätzung wird jedem Bin nun entsprechend der in ihm enthaltenen Daten ein Wert zugeteilt, der die Dichte in diesem Bin auf Grundlage der Gleichverteilungsannahme approximiert. Üblicher-

weise werden für die Bins (*d*-dimensionale) halboffene Intervalle gewählt, aber auch andere Borel-Mengen sind möglich.

Definition 2.14 (*allgemeiner Histogrammdichteschätzer*):

Sei $\{X_1,...,X_n\}$ eine einfache Zufallsstichprobe von n Elementen einer Verteilung mit Dichte f: $\Re^d \to \Re$. Desweiteren sei eine vollständige, disjunkte Partitionierung des Datenraums in $k < \infty$ Bereiche C_i, genannt Histogramm-Bins, gegeben. Dann ist der *allgemeine Histogrammdichteschätzer* $\hat{f}_H$ definiert als

$$\hat{f}_H(x) = \frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i}{Vol(C_i)} \cdot I_{C_i}(x) \tag{2.28}$$

mit $n_i = \left|\{X_j \in C_i, j = 1 \ldots n\}\right|$ und $I_{C_i}(x) = \begin{cases} 1 \text{ falls } x \in C_i \\ 0 \text{ sonst} \end{cases}$.

$Vol(C)$ bezeichne das Volumen von C.

Bemerkung 2.5:

- Eine andere Schreibweise für Gleichung (2.28) ist

$$\hat{f}_H(x) = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{1}{Vol(C_j)} \cdot I_{C_j}(X_i) I_{C_j}(x) \text{ mit } I_{C_j}(x) = \begin{cases} 1 \text{ falls } x \in C_j \\ 0 \text{ sonst} \end{cases} \tag{2.29}$$

- Handelt es sich bei den Histogramm-Bins um *d*-dimensionale Intervalle (achsenparallele (Hyper-)Rechtecke) der Form $C_i = (c_{i,min}, c_{i,max}]$, $c_i \in \Re^d$, so läßt sich (2.28) schreiben als

$$\hat{f}_{H,I}(x) = \frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i}{\prod_{j=1}^{d} h_{ij}} \cdot I_{C_i}(x) \text{ mit } h_{ij} = c_{i,max,j} - c_{i,min,j} \tag{2.30}$$

Ist außerdem die Dimension $d = 1$, so vereinfacht sich (2.30) noch weiter zu

$$\hat{f}_{H,1}(x) = \frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i}{h_i} \cdot I_{C_i}(x) \text{ mit } h_i = c_{i,max} - c_{i,min}. \tag{2.31}$$

Man beachte, daß in der Definition keine weiteren Aussagen über die Partitionierung getroffen wurden. Diese ist i.a. nicht eindeutig, so daß man abhängig von der Art der Partitionierung verschiedene Histogramm-Typen unterscheiden kann. Die in der Praxis gebräuchlichsten Typen sind das Equi-Width- und das Equi-Depth-Histogramm. Beim *Equi-Width-Histogramm* (EW)

haben alle Partitionen die gleiche Form, insbesondere das gleiche Volumen, d.h. $Vol(C_i) = const$ für alle i. Für achsenparallele (Hyper-)Rechtecke gilt genauer, daß die Seitenlängen der Kanten für jede Dimension bei allen Bins gleich ist: $h_{ij} = h_j$ für alle i. Dann lassen sich die Gleichungen (2.28) bis (2.31) schreiben als:

$$\text{EW-allgemein: } \hat{f}_{EW}(x) = \frac{1}{n \cdot Vol(C)} \cdot \sum_{i=0}^{k-1} n_i \cdot I_{C_i}(x),$$

$$\text{EW-Intervall: } \hat{f}_{EW,I}(x) = \frac{1}{n \cdot \prod_{j=1}^{d} h_j} \cdot \sum_{i=0}^{k-1} n_i \cdot I_{C_i}(x), \tag{2.32}$$

$$\text{EW-1-dim.: } \hat{f}_{EW,1}(x) = \frac{1}{n \cdot h} \cdot \sum_{i=0}^{k-1} n_i \cdot I_{C_i}(x).$$

Ein Beispiel für einen univariaten Equi-Width-Histogrammschätzer findet sich in Abbildung 2.4. Geschätzt wurde die Standardnormalverteilung im Bereich $[-4, 4]$ mit Mittelwert 0 und Standardabweichung 1 auf einer Stichprobe der Größe $n = 100$. Der Histogrammschätzer startet bei $x_0 = -4$ und besteht aus $k = 15$ Bins mit der Binweite $h = 8/15$. Die Säulengrafik zeigt die Histogrammbins C_i mit der jeweiligen in ihr enthaltenen Anzahl n_i der Stichprobenelemente. Die Kurve zeigt den Verlauf der Dichteschätzung (Sprünge sind der Einfachheit halber durchgezeichnet).

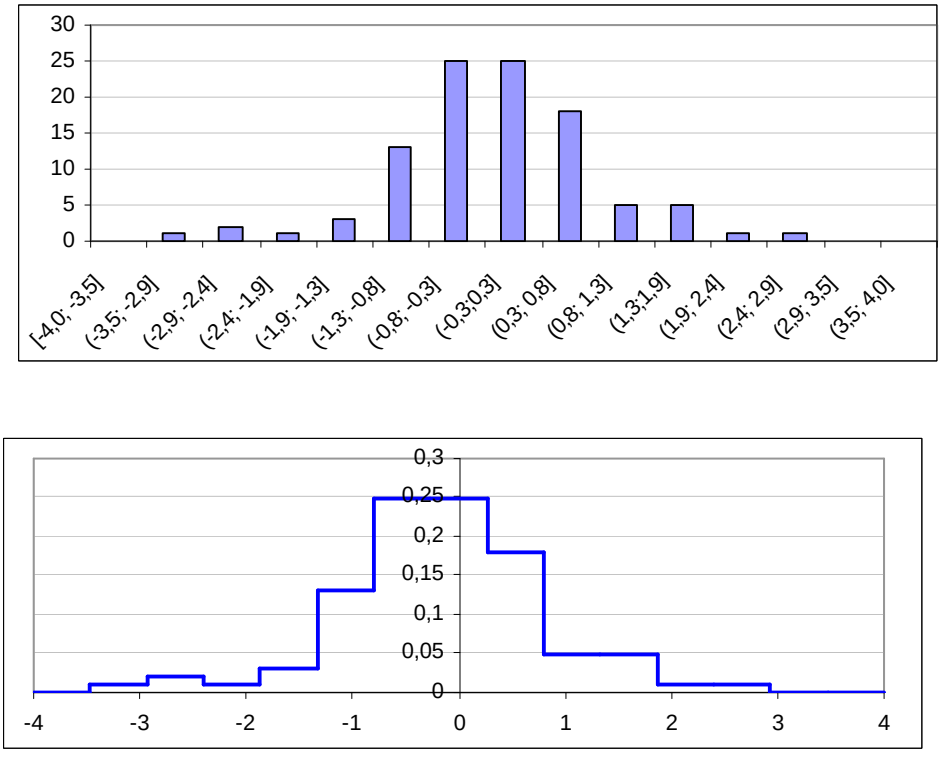


Abbildung 2.4: Beispiel für Equi-Width-Histogrammschätzer.

Beim *Equi-Depth-Histogramm* (ED) sind die Bins so gewählt, daß in jedes Bin gleich viele Elemente fallen, d.h. $n_i = n/k$ für alle i (oBdA. n/k ganzzahlig). Gleichungen (2.28) bis (2.31) vereinfachen sich dann zu:

$$\text{ED-allgemein: } \hat{f}_{ED}(x) = \frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{1}{Vol(C_i)} \cdot I_{C_i}(x),$$

$$\text{ED-Intervall: } \hat{f}_{ED,I}(x) = \frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{1}{\prod_{j=1}^{d} h_{ij}} \cdot I_{C_i}(x), \tag{2.33}$$

$$\text{ED-1-dim.: } \hat{f}_{ED,I}(x) = \frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{1}{h_i} \cdot I_{C_i}(x).$$

Ein Beispiel für einen univariaten Equi-Depth-Histogrammschätzer findet sich in Abbildung 2.5. Geschätzt wurde die Standardnormalverteilung im Bereich $[-4, 4]$ auf einer Stichprobe der Größe $n = 100$. Der Histogrammschätzer startet bei $x_0 = -4$ und besteht aus $k = 10$ Bins mit je 10 Elementen und variabler Binweite. Die Säulengrafik zeigt die Histogrammbins C_i mit der jeweiligen in ihr enthaltenen Anzahl n_i der Stichprobenelemente. Die Kurve zeigt den Verlauf der Dichteschätzung (Sprünge sind der Einfachheit halber durchgezeichnet).

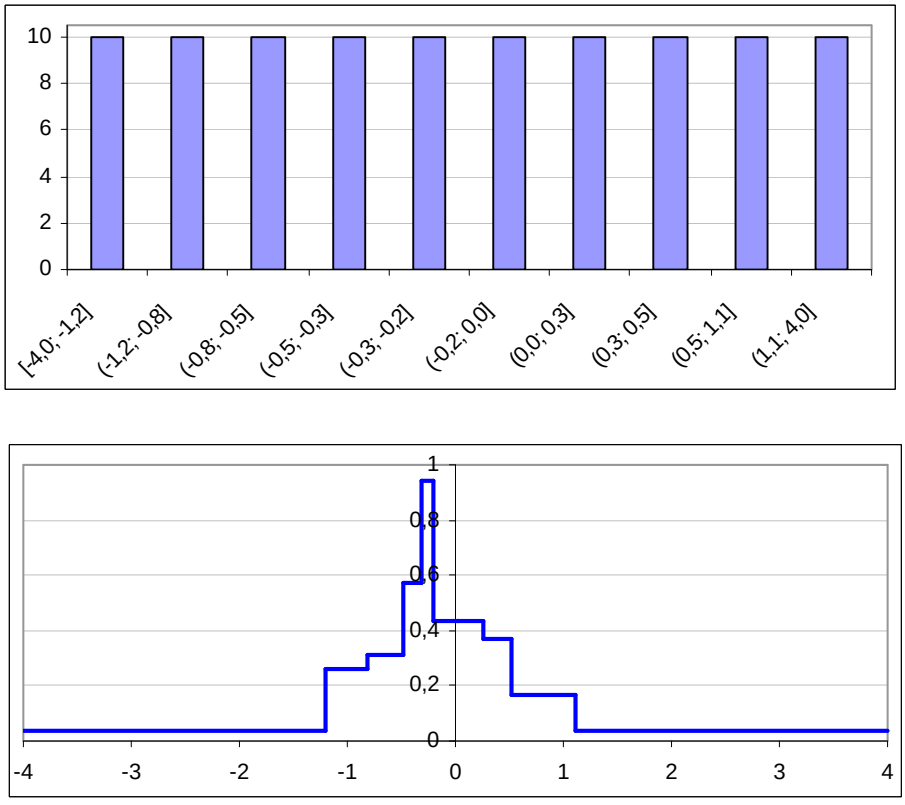


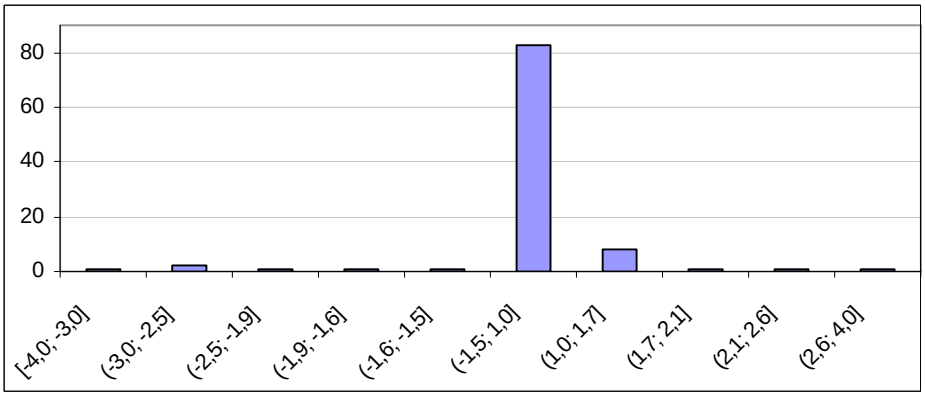
Abbildung 2.5: Beispiel für Equi-Depth-Histogrammschätzer.

Die vorgestellten Varianten seien noch einmal der Übersicht halber in der folgenden Tabelle zusammengefaßt. In den Spalten wird unterschieden ob es sich um beliebige Bins handelt oder um *d*-dimensionale Intervalle, mit $d = 1$ in der letzten Spalte. In den Zeilen wird unterschieden nach allgemeinen Histogrammen, Equi-Width-Histogrammen (EW) oder Equi-Depth-Histogrammen (ED):

	bel. Bins	***d*-dim. Intervalle**	**dim = 1**
allg.	$\frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i}{Vol(C_i)} \cdot I_{C_i}(x)$	$\frac{1}{n} \sum_{i=0}^{k-1} \frac{n_i}{\prod_{j=1}^{d} h_{ij}} I_{C_i}(x)$	$\frac{1}{n} \sum_{i=0}^{k-1} \frac{n_i}{h_i} I_{C_i}(x)$
EW	$\frac{1}{n \cdot Vol(C)} \sum_{i=0}^{k-1} n_i \cdot I_{C_i}(x)$	$\frac{1}{n \prod_{j=1}^{d} h_j} \sum_{i=0}^{k-1} n_i I_{C_i}(x)$	$\frac{1}{n h_i} \sum_{i=0}^{k-1} n_i I_{C_i}(x)$
ED	$\frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{1}{Vol(C_i)} \cdot I_{C_i}(x)$	$\frac{1}{k} \sum_{i=0}^{k-1} \frac{1}{\prod_{j=1}^{d} h_{ij}} I_{C_i}(x)$	$\frac{1}{k} \sum_{i=0}^{k-1} \frac{1}{h_i} I_{C_i}(x)$

Tabelle 2.1: Übersicht über verschiedene Histogramm-Dichteschätzer

Eine weitere Alternative ist das Max-Diff Histogramm, das von [Poosala et al. 96] für die eindimensionale Selektivitätsschätzung eingeführt wurde. Hier sind die Bins nach der größten Differenz benachbarter Stichprobenwerte partitioniert. Eine vereinfachte Formel läßt sich hierbei allerdings nicht angeben. Ein Beispiel für einen univariaten Max-Diff-Histogrammschätzer findet sich in Abbildung 2.6. Geschätzt wurde die Standardnormalverteilung im Bereich $[-4, 4]$ auf einer Stichprobe der Größe $n = 100$. Der Histogrammschätzer startet bei $x_0 = -4$ und besteht aus $k = 10$ Bins mit variabler Anzahl Elemente und variabler Binweite. Die Säulengrafik zeigt die Histogrammbins C_i mit der jeweiligen in ihr enthaltenen Anzahl n_i der Stichprobenelemente. Die Kurve zeigt den Verlauf der Dichteschätzung (Sprünge sind der Einfachheit halber durchgezeichnet).



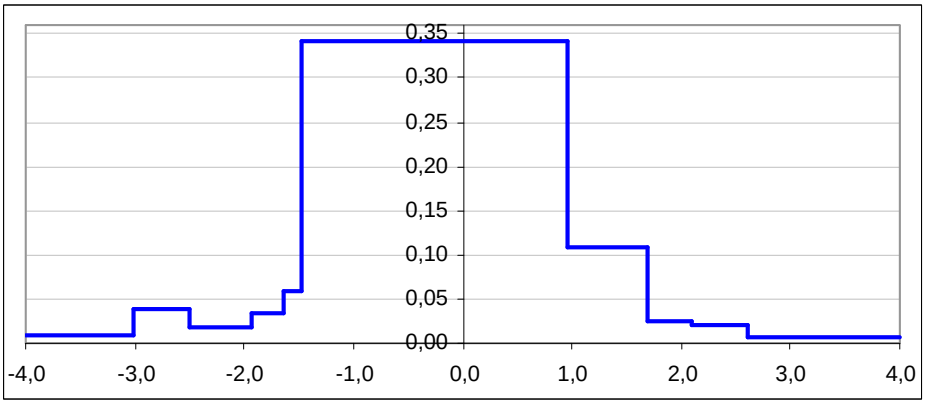


Abbildung 2.6: Beispiel für Max-Diff-Histogrammschätzer.

Andere mehrdimensionale Histogramme zur Selektivitätsschätzung verwenden Partitionierungen in Analogie zu mehrdimensionalen Indexstrukturen ([Samet 90], [Seeger 96]).

Obwohl Histogramme bereits zahlreich in der Literatur als Methode zur Selektivitätsschätzung diskutiert wurde, fehlte es bisher an einer korrekten mathematischen Definition. In Kapitel 3.2 werden daher Histogrammselektivitätsschätzer explizit definiert.

Im Gegensatz zu der empirischen Verteilungsfunktion sind die Histogrammschätzer i.a. nicht erwartungstreu, vgl. [Scott 92], [Büning & Trenkler 94]. Einfache statistische Eigenschaften des univariaten Equi-Width-Histogrammschätzers leiten sich aus der Tatsache ab, daß die n_j, die Anzahl der Elemente in C_j, binomiale Zufallszahlen sind. Dann gilt mit $p_j = \int_{C_j} f(t)dt$:

$$E(n_j) = np_j \text{ für } x \in C_j \text{ und } Var(n_j) = np_j(1-p_j) \tag{2.34}$$

Daraus folgt für den Erwartungswert des univariaten Histogrammschätzers

$$E(\hat{f}_{EW,1}(x)) = \frac{E(n_j)}{nh} = \frac{p_j}{h} \text{ für } x \in C_j \tag{2.35}$$

und für die Varianz des univariaten Histogrammschätzers

$$Var(\hat{f}_{EW,1}(x)) = \frac{Var(n_j)}{(nh)^2} = \frac{p_j(1-p_j)}{nh^2} \text{ für } x \in C_j \tag{2.36}$$

Der Histogrammschätzer ist also nur dann erwartungstreu, wenn die Dichte $f(x)$ konstant über dem Intervall C_j mit $x \in C_j$, d.h. stückweise gleichverteilt ist.

Der Histogrammschätzer besitzt zudem noch zwei weitere Nachteile: Zum einen ist das (Equi-Width-) Histogramm und damit die Güte des Schätzers abhängig von der Wahl des Startpunktes

der Partition (vgl. hierzu z.B. [Scott 92], [Gasser & Müller 79], [Wand & Jones 95]). Zum anderen handelt es sich bei der Schätzfunktion um eine i.a. nicht analytisch stetige Treppenfunktion, wogegen die zugrunde liegende Dichtefunktion durchaus stetig sein kann. Es existieren nun zwei Weiterentwicklungen des Histogrammschätzers, die jeweils eine der beiden zuletzt genannten Nachteile vermeiden. Diese seien im folgenden Abschnitt vorgestellt.

2.3.3 Weiterentwicklungen des Histogrammschätzers

In diesem Abschnitt werden zwei Weiterentwicklungen des Histogrammschätzers vorgestellt, die jeweils eine der oben genannten Nachteile vermeiden. Bei den Häufigkeitspolygonschätzern wird eine analytisch stetige Funktion gebildet, während bei den Average Shifted Histogrammschätzern die Abhängigkeit vom Startpunkt vermindert wird.

Häufigkeitspolygonschätzer

Bei den Häufigkeitspolygonen werden die Zentren benachbarter Histogramm-Bins in der Höhe der lokalen Dichteschätzung miteinander verbunden, so daß sich eine analytisch stetige, aber i.a. nicht differenzierbare Schätzfunktion ergibt.

Sei beispielsweise ein univariates Equi-Width-Histogramm $\hat{f}_H$) gegeben. Liege weiter x im Intervall $(-h/2, h/2)$. Dann ist der zugehörige Häufigkeitspolygonschätzer $\hat{f}_{FP}$ gegeben durch

$$\hat{f}_{FP}(x) = \hat{f}_H\left(-\frac{h}{2}\right)\left(\frac{1}{2} - \frac{x}{h}\right) + \hat{f}_H\left(\frac{h}{2}\right)\left(\frac{1}{2} + \frac{x}{h}\right) \text{ für } x \in \left(-\frac{h}{2}, \frac{h}{2}\right) \tag{2.37}$$

Ein Beispiel für einen univariaten *Häufigkeitspolygonschätzer* findet sich in Abbildung 2.7. Geschätzt wurde die Standardnormalverteilung im Bereich $[-4, 4]$ auf einer Stichprobe der Größe $n = 100$. Der ursprüngliche Equi-Width-Histogrammschätzer startet bei $x_0 = -4$ und besteht aus $k = 15$ Bins der Binweite $h = 8/15$.

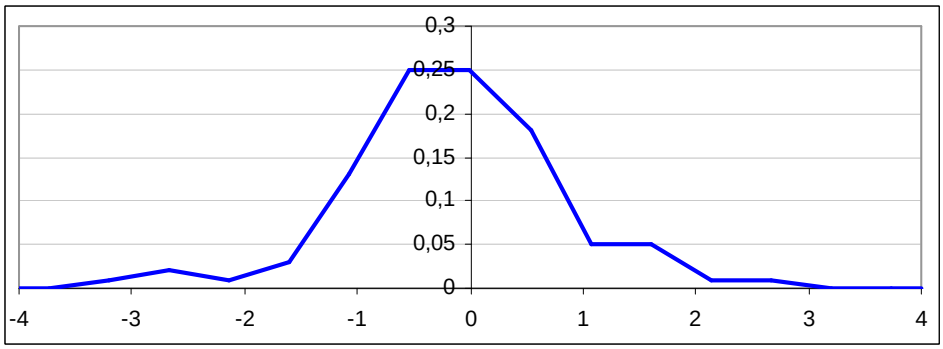


Abbildung 2.7: Beispiel für Häufigkeitspolygonschätzer.

Mögliche Erweiterungen des multivariaten Häufigkeitspolygonschätzers bestehen darin, die Zentren der hyper-rechteckigen Histogramm-Bins auf geeignete Weise zu verbinden - dies ist jedoch i.a. nicht eindeutig, vgl. hierzu z.B. [Scott 92], S. 106.

Bemerkung 2.6:

Ebenso wie der Histogrammschätzer ist der Häufigkeitspolygonschätzer nicht erwartungstreu.

Da der Häufigkeitspolygonschätzer immer stetig ist, können mögliche Sprungstellen der wahren Dichte nicht abgebildet werden. Dies ist ein Nachteil gegenüber den Histogrammschätzern, wo die Bins so gelegt werden können, daß die Grenze zwischen zwei benachbarten Bins genau auf so eine Sprungstelle fällt - vorausgesetzt die Sprungstelle ist bekannt oder kann durch geeignete Verfahren ermittelt werden. Dieses Problem taucht auch bei den im nächsten Abschnitt vorgestellten Kernschätzern auf und wird in Kapitel 3.5 im Zusammenhang mit der Selektivitätsschätzung behandelt.

Da die im nächsten Abschnitt vorgestellten Kernschätzer außerdem die Eigenschaft der Differenzierbarkeit haben und nicht die Konstruktion von Histogrammen erfordern, wird der Ansatz der Häufigkeitspolygonschätzer in dieser Arbeit nicht weiter verfolgt.

Average Shifted Histogrammschätzer

Bei den *Average Shifted Histogrammen* werden mehrere (Equi-Width-) Histogramme mit gleicher Anzahl von Bins aber unterschiedlicher Startposition überlagert. Das Equi-Width-Histogramm wird dabei - für jede Dimension - um entsprechende Anteile (*Shifts*) in Richtung der entsprechenden Dimension verschoben. Die Länge der Shifts in jeder Dimension *i* ist gleich der Binweite h_i in der jeweiligen Dimension geteilt durch die Anzahl m_i der Shifts in der jeweiligen Dimension. Auf diese Weise entfällt die starke Abhängigkeit des Schätzers von der Startposition mit der Anzahl der Shifts m_i. Für den univariaten Fall ist der Average Shifted Histogrammdichteschätzer wie folgt definiert.

Definition 2.15 (*univariater Average Shifted Histogrammdichteschätzer, ASH*):

Sei eine einfache Zufallsstichprobe $X_1,...,X_n$ von n Elementen einer Verteilung mit Dichte f: $\Re \rightarrow \Re$ sowie eine Menge von m Equi-Width-Histogrammschätzern $\hat{f}_1,...,\hat{f}_m$ mit gleicher Binweite h aber unterschiedlichen Startpunkten $x_0 = 0, h/m, ..., (m-1)h/m$ gegeben. Dann ist der *Average Shifted Histogramm Schätzer* mit k Bins und m Shifts definiert als:

$$\hat{f}_{ASH}(x) = \frac{1}{m} \sum_{j=1}^{m} \hat{f}_j(x) \tag{2.38}$$

Im multivariaten Fall werden die Equi-Width-Histogramme in jede Dimension um entsprechende Shifts verschoben. Die Anzahl m_i der Shifts und die Binweite h_i kann dabei für jede

Dimension variieren. Es werden somit $m_1 \times \ldots \times m_d$ Equi-Width-Histogramme überlagert. Der multivariate Average Shifted Histogramm Schätzer berechnet sich somit gemäß:

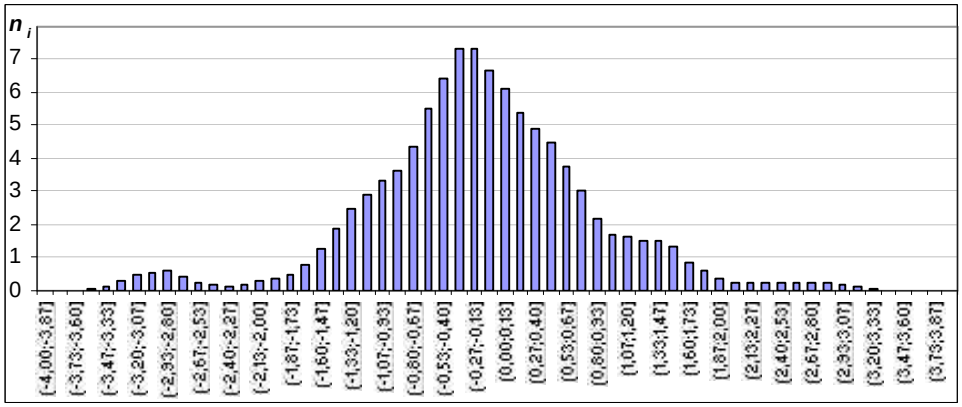
$$\hat{f}_{ASH}(x) = \frac{1}{m_1 \ldots m_d} \sum_{j_1 = 1}^{m_1} \ldots \sum_{j_d = 1}^{m_d} \hat{f}_{j_1 \ldots j_d}(x) \tag{2.39}$$

Dabei ist $\hat{f}_{j_1 \ldots j_d}(x)$ dasjenige *d*-dimensionale Histogramm mit dem Startpunkt $((j_1 - 1)h_{j_1} / m_{j_1}, \ldots, (j_d - 1)h_{j_d} / m_{j_d})^t$.

Bemerkung 2.7:

> Es läßt sich zeigen, daß bei geeigneter Definition des ASH-Schätzers dieser im Grenzfall, d.h. für $m \to \infty$, äquivalent ist zu den im folgenden vorgestellten Kernschätzern, vgl. hierzu z.B. [Scott 92], S. 121f.

Ein Beispiel für einen univariaten Average Shifted Histogrammschätzer findet sich in Abbildung 2.6. Geschätzt wurde die Standardnormalverteilung im Bereich $[-4, 4]$ auf einer Stichprobe der Größe $n = 100$. Der Average Shifted Histogrammschätzer startet bei $x_0 = -4$ und alle vier Shifts bestehen jeweils aus $k = 16$ Bins mit variabler Anzahl Elemente und fester Binweite. Die Säulengrafik zeigt die Average Shifted Histogramm-Bins mit der jeweiligen in ihr enthaltenen Anzahl n_i der Stichprobenelemente. Die Kurve zeigt den Verlauf der Dichteschätzung (Sprünge sind der Einfachheit halber durchgezeichnet).



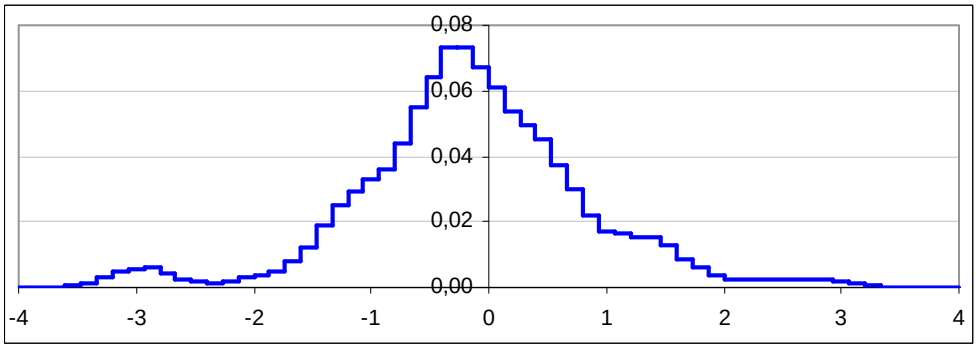


Abbildung 2.8: Beispiel für Average Shifted Histogrammschätzer mit 4 Shifts und jeweils 16 Bins.

Average Shifted Histogramme werden in Kapitel 3.3 ebenfalls zur Selektivitätsschätzung erweitert und in Kapitel 5 sowohl im univariaten als auch im bivariaten Fall untersucht und mit den anderen Schätzverfahren verglichen.

2.3.4 Kerndichteschätzer

Wie bereits in vorigen Abschnitten erwähnt, sind die Nachteile von Histogrammschätzern, daß sie i.a. nicht erwartungstreu, i.a. keine glatte Schätzung liefern und abhängig von der Wahl des Startpunktes sind. Erste Lösungsansätze um diese Nachteile zu eliminieren sind die im vorherigen Abschnitt beschriebenen Erweiterungen der Histogrammschätzer. Diese sind aber nur zum Teil befriedigend. Gesucht ist vielmehr eine glatte, stetig differenzierbare Funktion mit Integralwert Eins, die die lokalen Gegebenheiten der zu schätzenden Dichtefunktion widerspiegelt. Ein allgemeiner Ansatz für eine solche Schätzfunktion besteht darin, den Durchschnitt einer geeigneten von der Stichprobe abhängigen Gewichtsfunktion *w* zu nehmen der Form (vgl. [Silverman 86]):

$$\hat{f}_G(x) = \frac{1}{n}\sum_{i=1}^{n} w(x, X_i) \tag{2.40}$$

Unter den Bedingungen $w(x, y) \geq 0 \quad \forall x, y \in \Re^d$ und $\int_{\Re^d} w(x, y)dx = 1$ ist sichergestellt, daß $\hat{f}_G$ die Eigenschaften einer Dichtefunktion besitzt.

Kernfunktionen ([Silverman 86], [Scott 92], [Wand & Jones 95]) stellen eine geeignete und elegante Alternative für die zu wählende Gewichtsfunktion dar. Dabei handelt es sich anschaulich um "Hügel" um die einzelnen Stichprobenelemente, die abhängig von einem zu bestimmenden Glättungsparameter in der Gleichung (2.40) überlagert werden und so die Schätzfunktion bilden. Abbildung 2.9 zeigt ein simples Beispiel mit 6 Stichprobenwerten. Über diesen sind als gestrichelte Linie die entsprechenden Kernfunktionen gezeichnet. Im Beispiel wurde der Epanechnikow-Kern (s.u.) gewählt. Die Kernfunktionen überlagern sich zur geschätzten Dichte-

funktion, die in der Abbildung durch eine durchgezogene Linie repräsentiert wird. Auf diese Art wird die "Wahrscheinlichkeitsmasse" eines einzelnen Stichprobenwertes auf seine Umgebung verteilt, so daß bei der Schätzung in einem Punkt nicht nur ein einzelner Stichprobenwert sondern auch dessen Nachbarn - je nach Bandbreite (s.u.) - berücksichtigt werden.

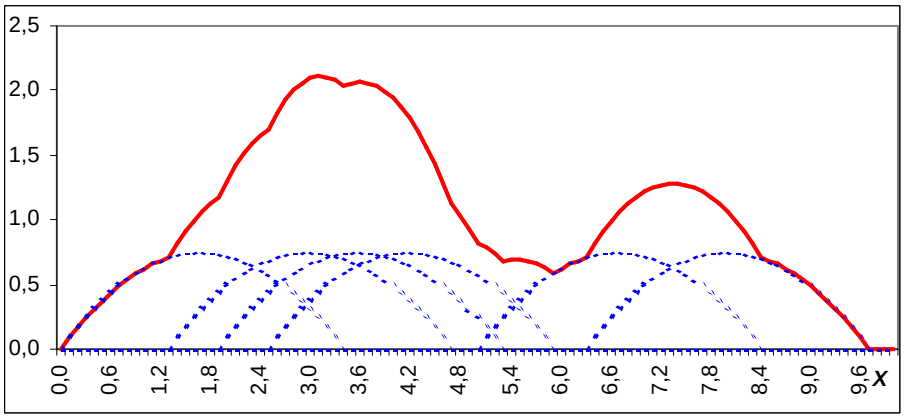


Abbildung 2.9: Überlagerung mehrer Kernfunktionen in Abhängigkeit von den Stichprobenwerten.

Kernfunktionen stellen eine ganze Klasse von Funktionen dar, die alle von einem Bandbreiten-Parameter abhängen. Die Wahl der speziellen Kernfunktion bestimmt das Aussehen der Schätzfunktion, während die Bandbreite den Einflußbereich der Kernfunktion festlegt und damit entscheidenden Einfluß auf das Glättungsverhalten des Schätzers besitzt.

Üblicherweise werden Kernfunktionen mit den Eigenschaften gewählt, daß sie uni-modal, symmetrisch bzgl. der Ordinatenachse und nicht-negativ sind und das unbestimmte Integral gleich 1 ist. Diese Eigenschaften sind aber nicht zwingend notwendig, vgl. z.B. [Granovsky et al. 95]. Auch werden in Kapitel 3.4.2 dieser Arbeit sogenannte Randkerne benutzt, die diese Eigenschaften nicht erfüllen.

Definition 2.16 (*Kernfunktion*):

Eine Funktion $K : \Re^d \to \Re$ heiße *Kernfunktion*, falls K beschränkt und Borel-meßbar ist

mit: $\int_{\Re^d} |K(t)|dt < \infty$ und $|K(t)| \to 0$ für $|t| \to \infty$. (Vgl. auch [Silverman 86], S. 71)

K heiße *normierte Kernfunktion*, falls es zudem die folgenden Eigenschaften erfüllt:

- Das Integral ist gleich Eins: $\int_{\Re^d} K(t)dt = 1$,
- K ist unimodal und symmetrisch bzgl. der Ordinatenachse,
- K ist nicht-negativ: $K(x) \geq 0 \forall x \in \Re^d$

Eine normierte Kernfunktion hat somit insbesondere die Eigenschaften einer Dichtefunktion.

Im folgenden seien einige einfache (normierte) univariate Kernfunktionen beschrieben.

Beispiel 2.1:

- Der *Rechteck-Kern* K_R ist definiert als (vgl. Abb. 2.10a)

$$K_R(t) = \begin{cases} 1/2 & \text{für } |t| \le 1 \\ 0 & \text{sonst} \end{cases}, \; t \in \Re \tag{2.41}$$

- Der *Dreieck-Kern* K_D ist definiert als (vgl. Abb. 2.10b)

$$K_D(t) = \begin{cases} 1 - |t| & \text{für } |t| \le 1 \\ 0 & \text{sonst} \end{cases}, \; t \in \Re \tag{2.42}$$

- Der *Gauß-Kern* K_{Gauss} ist definiert als (vgl. Abb. 2.10c)

$$K_{Gauss}(t) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-t^2}{2}\right), \; t \in \Re \tag{2.43}$$

a) **b)** **c)**

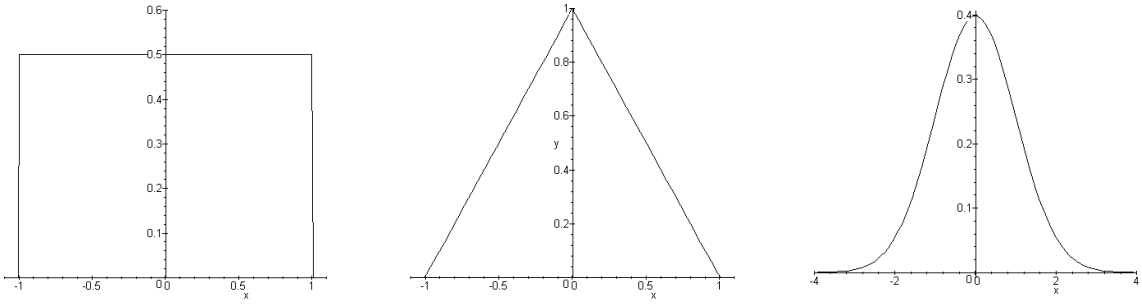


Abbildung 2.10: a) Rechteck-Kern, b) Dreieck-Kern, c) Gauß-Kern

- Der *Epanechnikow-Kern* K_{Epa} ist definiert als (vgl. Abb. 2.11a)

$$K_{Epa}(t) = \begin{cases} \frac{3}{4}(1 - t^2) & \text{für } |t| \le 1 \\ 0 & \text{sonst} \end{cases}, \; t \in \Re \tag{2.44}$$

- Der *Biweight-Kern* K_{Biw} ist definiert als (vgl. Abb. 2.11b)

$$K_{Biw}(t) = \begin{cases} \frac{15}{16}(1 - t^2)^2 & \text{für } |t| \le 1 \\ 0 & \text{sonst} \end{cases}, \; t \in \Re \tag{2.45}$$

a) **b)**

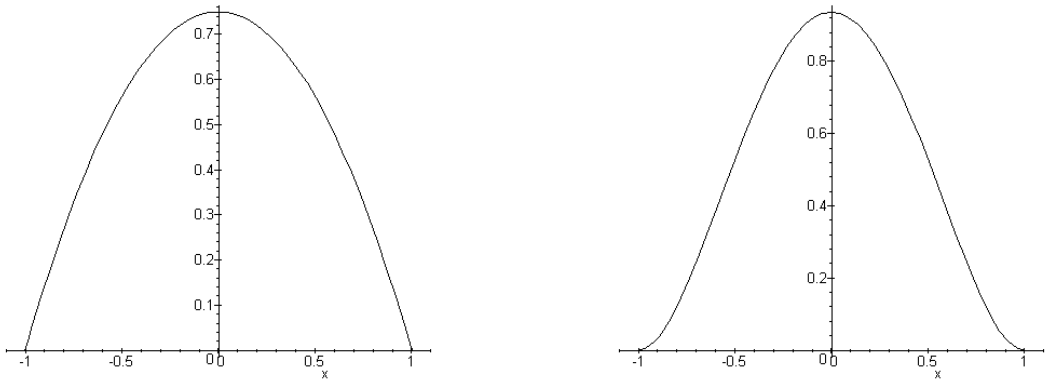


Abbildung 2.11: a) Epanechnikow-Kern, b) Biweight-Kern

Alle fünf aufgeführten Kernfunktionen haben die Eigenschaft einer unimodalen bzgl. der Ordinatenachse symmetrischen und nicht-negativen Dichtefunktion mit $\int_{-\infty}^{\infty} K(t)dt = 1$. Die Eigenschaft der Dichtefunktion ist insbesondere sinnvoll im Hinblick auf die Schätzung einer Dichtefunktion mittels Kernfunktionen. Bis auf den Gauß-Kern, der unendlichen Träger hat, haben diese Kernfunktionen den Träger $[-1,1]$. Der Epanechnikow-Kern und der Biweight-Kern sind stetig differenzierbar, der Gauß-Kern sogar beliebig oft. Aufgrund der einfacheren Berechenbarkeit, insbesondere mit Computern, sind in praktischen Anwendungen der Epanechnikow- oder der Biweight-Kern vorzuziehen.

Bisher wurden lediglich Beispiele für univariate Kernfunktionen vorgestellt (Gleichungen (2.41) bis (2.45)). Es existieren nun zwei gängige Verfahren zur Herleitung von multivariaten Kernfunktionen aus univariaten Kernfunktionen. Dabei werden Produktkerne und sphärische Kerne unterschieden - die Namen ergeben sich aus der Art der Herleitung:

Definition 2.17 (*Produktkern, Sphärischer Kern*):

Sei K: $\Re \to \Re$ eine univariate bzgl. der Ordinatenachse symmetrische Kernfunktion.

Dann ist der d-variate *Produktkern* K^P definiert als:

$$K^P: \Re^d \to \Re,\ K^P(x) = \prod_{j=1}^{d} K(x_j), \tag{2.46}$$

und der d-variate *sphärische Kern* K^S ist definiert als:

$$K^S: \Re^d \to \Re,\ K^S(x) = c_{K,d} \cdot K\left((x^t x)^{1/2}\right) \tag{2.47}$$

mit $c_{K,d} = \int_{\Re^d} K\left((x^t x)^{1/2}\right)dx$ konstant.

Beispiel 2.2:

- Durch Einsetzen des Gauß-Kerns K_{Gauss} (2.43) entweder in Gleichung (2.46) oder in Gleichung (2.47) ergibt sich in beiden Fällen als *multivariater Gauß-Kern*:

$$K^S_{Gauss}(x) = K^P_{Gauss}(x) = \left(\frac{1}{2\pi}\right)^{d/2} \cdot \exp(-\frac{x^t x}{2}) \tag{2.48}$$

  Beim Gauß-Kern führen beide Ansätze zum selben Ergebnis. Dies gilt i.a. nicht für die anderen Kernfunktionen.

- Durch Einsetzen des Epanechnikow-Kerns K_{Epa} (2.44) in Gleichung (2.46) ergibt sich als *Epanechnikow-Produktkern* mit $x = (x_1,\ldots, x_d)^t \in \Re^d$

$$K^P_{Epa}(x) = \left(\frac{3}{4}\right)^d \cdot \prod_{j=1}^{d} (1 - x_j^2) \cdot 1_{-1 \le x_j \le 1} \tag{2.49}$$

- Der bivariate Epanechnikow-Produktkern mit $x = (x_1,x_2)^t \in \Re^2$ lautet somit:

$$K^P_{Epa}(x) = \begin{cases} \frac{9}{16}(1 - x_1^2)(1 - x_2^2) & \text{, falls } |x_1|, |x_2| \le 1 \\ 0 & \text{, sonst} \end{cases}, \tag{2.50}$$

  vgl. Abbildung 2.12.

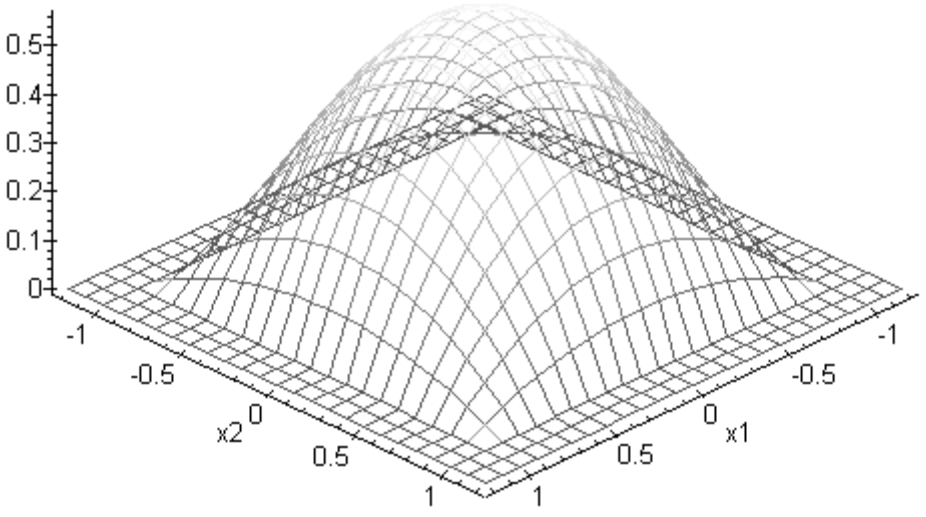


Abbildung 2.12: bivariater Epanechnikow-Produktkern

Für eine Diskussion der beiden unterschiedlichen Ansätze zur Verwendung von Kernfunktionen zur Dichteschätzung vgl. [Wand & Jones 95], Kap. 4.5, S. 103 f. Dort wird auch gezeigt, daß der sphärische Kern bzgl. des AMISE leicht besser ist. Für praktische Anwendungen ist jedoch wegen der einfacheren Berechenbarkeit im multivariaten Fall der Produktkernschätzer zu empfehlen. Dieser wird in Kapitel 3.4 zur multivariaten Selektivitätsschätzung mit Kernschätzern verwendet.

Um den Einfluß der Wahl einer Kernfunktion auf einen Schätzer beurteilen zu können, wird der Begriff der Ordnung einer Kernfunktion benutzt

Definition 2.18 (*Ordnung einer Kernfunktion*):

Die *Momente* $\kappa_i = \kappa_i(K)$ einer univariaten Kernfunktion K, $i \geq 0$, seien definiert als

$$\kappa_i := \int_{\Re} t^i K(t)dt \tag{2.51}$$

Eine Kernfunktion K sei von der *Ordnung* $p > 1$, falls gilt:

$$\begin{aligned} &\text{a) } \kappa_0 = 1, \\ &\text{b) } \kappa_i = 0 \ \forall i = 1, \ldots, p-1 \text{ und} \\ &\text{c) } \kappa_p \neq 0 \end{aligned} \tag{2.52}$$

In Kapitel 4 sind im multivariaten Fall Kernfunktionen 2. Ordnung von Interesse.

Eine multivariate Kernfunktion K sei von der *Ordnung* 2, falls gilt:

$$\begin{aligned} &\text{a) } \kappa_0 := \int_{\Re^d} K(\boldsymbol{t})d\boldsymbol{t} = \boldsymbol{1}, \\ &\text{b) } \kappa_1 := \int_{\Re^d} \boldsymbol{t}K(\boldsymbol{t})d\boldsymbol{t} = \boldsymbol{0} \text{ und} \\ &\text{c) } \kappa_2 := \int_{\Re^d} \boldsymbol{t}\boldsymbol{t}^T K(\boldsymbol{t})d\boldsymbol{t} = \boldsymbol{I}_d \end{aligned} \tag{2.53}$$

Mit Hilfe der Ordnung von Kernen ist es möglich genauere Aussagen über die Optimalität von Kernen zu treffen, s. unten.

Um nun Kernfunktionen zur Dichteschätzung zu verwenden, wird im folgenden - wenn nicht anders gesagt - vorausgesetzt, daß es sich bei der Kernfunktion K um eine normierte Kernfunktion im Sinne von Definition 2.16 handelt.

Betrachtet man den allgemeinen Gewichtsfunktionsschätzer aus Gleichung (2.40), so erhält man einen Schätzer unter Verwendung von Kernfunktionen, indem man $w(x, y) = K_H(x - y)$

setzt mit $K_H(x) = |H|^{-1}K(H^{-1}x)$ und H einer nicht-singulären $d \times d$-Matrix, der sog. Bandbreitenmatrix. Dabei sind die Voraussetzungen $K_H(x-y) \geq 0$ und $\int_{\Re^d} K_H(x-y)dx = \int_{\Re^d} K(t)dt = 1$ erfüllt. Somit läßt sich der Kerndichteschätzer wie folgt definieren:

Definition 2.19 (*allgemeiner Kerndichteschätzer*):

Sei $\{X_1,...,X_n\}$ eine einfache Zufallsstichprobe von n Elementen einer Verteilung mit stetiger Dichte f: $\Re^d \rightarrow \Re$. Sei K eine d-variate normalisierte Kernfunktion. Sei $H = H(x,X_i)$ eine nicht-singuläre $d \times d$-Matrix - H heiße *Bandbreitenmatrix*. Dann ist der *allgemeine Kerndichteschätzer* der Dichte f definiert als:

$$\hat{f}_K(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} K_H(x - X_i) \text{ mit } K_H(x) = \frac{1}{|H|}K(H^{-1}x) \tag{2.54}$$

Bemerkung 2.8:

- Im univariaten Fall schreibt sich Gleichung (2.54) mit der Bandbreite $h = h(x,X_i) > 0$ als

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}K\left(\frac{x - X_i}{h}\right). \tag{2.55}$$

  Ist h unabhängig von den X_i, so läßt sich der Faktor $1/h$ aus der Summe herausziehen.

- Durch die Wahl spezieller Bandbreitenmatrizen H läßt sich auch für multivariate Kerndichteschätzer Gleichung (2.54) vereinfachen. Eine Möglichkeit besteht darin als Bandbreitenmatrix eine Diagonalmatrix $H = \text{diag}(h_1,...h_d)$ mit $h_j \in \Re_+$ zu wählen. Dann vereinfacht sich (2.54) zu:

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\tilde{h}} \cdot K\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_d}\right) \text{ mit } \tilde{h} = \prod_{j=1}^{d} h_j \tag{2.56}$$

  Die Glättung wird hiermit über die Randverteilungen gesteuert. Dies ist eine in vielen praktischen Fällen zu empfehlende Vereinfachung. Notfalls ist zuvor eine Hauptachsentransformation [Scott 92] durchzuführen.

  Eine weitere Möglichkeit besteht darin, in der obigen Diagonalmatrix alle h_j gleich einem h zu wählen, so daß sich aus Gleichung (2.56) ergibt

$$\hat{f}_K(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d}K\left(\frac{1}{h}(x - X_i)\right) \tag{2.57}$$

Dies ist eine weitere Vereinfachung, die jedoch in der Praxis i.a. nicht realistisch ist, da hier davon ausgegangen wird, daß die Daten bzgl. aller Randverteilungen gleiche Varianz und Bias besitzen (siehe hierzu Kapitel 4).

In dieser Arbeit wird daher im folgenden und insbesondere zur Selektivitätsschätzung in Kapitel 3.4 im multivariaten Fall der Kernschätzer mit Diagonalmatrix $H = \text{diag}(h_1^2,...h_d^2)$ mit $h_j \in \Re^{>0}$ verwendet.

Bemerkung 2.9:

I.a. wird die Kernfunktion *K* so gewählt, daß es sich um eine unimodale bzgl. der Ordinatenachse symmetrische Dichtefunktion handelt. Dies gewährleistet, daß es sich auch bei der Funktion $\hat{f}_K$ um eine Dichtefunktion handelt. Es gibt allerdings Fälle, in denen auch andere Kernfunktionen zum Tragen kommen. Ein solcher Fall wird in Abschnitt 2.3.5 im Zusammenhang mit Randkernen vorgestellt.

Ein Beispiel für einen univariaten Kerndichteschätzer findet sich in Abbildung 2.13. Geschätzt wurde die Standardnormalverteilung im Bereich [–4 , 4] auf einer Stichprobe mit *n* = 50 Elementen. Der Kernschätzer wurde entsprechend den Kurven mit drei verschiedenen jeweils festen Bandbreiten *h* berechnet und an 80 Stellen ausgewertet.

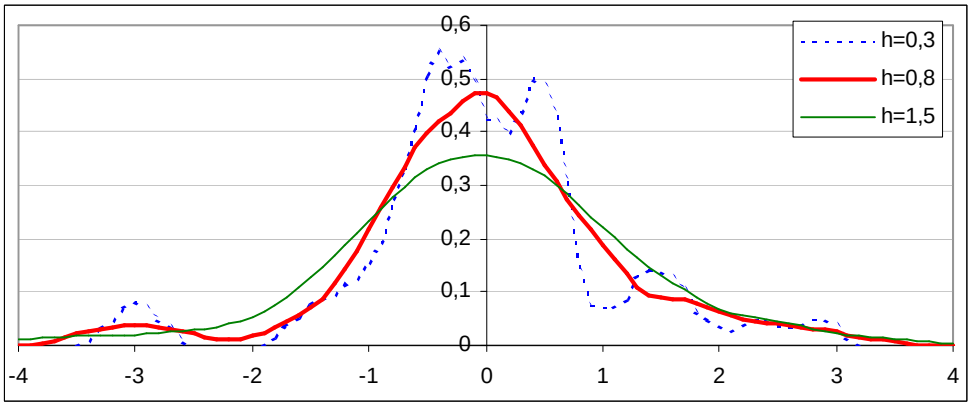


Abbildung 2.13: Beispiel für Kerndichteschätzer.

Bei der Benutzung von Kernschätzern gibt es drei unbekannte Größen, einmal die Wahl der Kernfunktion, zum zweiten die Bestimmung der Bandbreitenmatrix *H* bzw. der Bandbreite *h* und natürlich die Stichprobe der Größe *n*. Sowohl mathematische Überlegungen als auch Experimente zeigen, daß dabei die kritischen Parameter die Wahl der Bandbreite und die Stichprobengröße sind, vgl. auch Kapitel 4 und 5.

Die Wahl der Kernfunktion hat keinen so großen Einfluß auf die Güte der Schätzung, vgl. hierzu z.B. [Scott 92], [Wand & Jones 95]. Nichtsdestotrotz unterscheiden sich die Kernfunktionen. So ist z.B. der Rechteck-Kern nicht stetig und der Dreieckskern nicht stetig differenzierbar. Man

kann weiterhin zeigen, daß der Epanechnikow-Kern optimal bzgl. der Minimierung des AMISE ist ([Scott 92], S. 138f.; [Wand & Jones 95], S. 28f. und S. 104 f.). Zudem ist der Epanechnikow-Kern - im Vergleich zum Gauß-Kern - leicht und effizient auszuwerten, insbesondere bei Computerberechnungen. Im folgenden wird daher im wesentlichen der Epanechnikow-Kern verwendet.

Der Bandbreiten-Parameter beeinflußt die Glattheit der Schätzung. Ist der Parameter zu groß gewählt, so werden detaillierte Strukturen in der wahren Dichtefunktionen verwischt, d.h. der Schätzer überglättet. Ist der Parameter dagegen zu klein gewählt, so paßt sich die Schätzfunktion zu stark den Stichprobenwerten an mit der Folge, daß Artefakte in der Schätzung der wahren Dichtefunktion entstehen können, vgl. Abbildung 2.13. Die Wahl der Bandbreite hat daher einen ähnlichen Einfluß auf die Güte der Schätzfunktionen wie die Anzahl der Bins beim Histogrammschätzer. Da der Bandbreiten-Parameter einen dermaßen starken Einfluß auf die Qualität der Schätzung hat, wird seiner Wahl ein eigenes Kapitel gewidmet, siehe Kapitel 4.

Da der Kernschätzer als arithmetisches Mittel von n unabhängigen und gleichverteilten Zufallsvariablen $K_h(x,X)$ angesehen werden kann, gilt für den Erwartungswert und die Varianz des Kernschätzers:

$$E(\hat{f}_K(x)) = E(K_h(x, X)) \text{ und } Var(\hat{f}_K(x)) = \frac{1}{n} \cdot Var(K_h(x, X)) \tag{2.58}$$

Aussagen in der Literatur über statistische Eigenschaften der Kernschätzer beziehen sich fast ausschließlich auf große Stichprobenumfänge und sind daher überwiegend asymptotische Aussagen. Unter den folgenden Annahmen (vgl. [Büning & Trenkler 94])

$$\int_{-\infty}^{\infty} |K(x)|dx < \infty \tag{2.59}$$

$$\sup(|K(x)|, x \in \Re) < \infty \tag{2.60}$$

$$\lim_{x \to \infty} |xK(x)| = 0 \tag{2.61}$$

ist der univariate Kernschätzer $\hat{f}(x) = \hat{f}_n(x)$ asymptotisch erwartungstreu, d.h.

$$\lim_{n \to \infty} E(\hat{f}_n(x)) = f(x) \tag{2.62}$$

Wie bereits zuvor dargestellt ist ein wichtiges Kriterium für die Güte eines Dichteschätzers die Konsistenz. Gelten weiterhin die beiden folgenden Bedingungen an die Bandbreite des Kernschätzers

$$\lim_{n \to \infty} h = 0 \text{ und } \lim_{n \to \infty} nh = \infty \text{ für } h = h_n, \tag{2.63}$$

so ist der univariate Kernschätzer konsistent im quadratischen Mittel, vgl. z.B. [Büning & Trenkler 94].

Der Begriff der Konsistenz spielt ebenfalls eine wichtige Rolle bei der Betrachtung von Randproblemen im nächsten Abschnitt.

2.3.5 Randprobleme

Aus der Theorie und Praxis der Kerndichteschätzer ist bekannt (z.B. [Silverman 86], [Jones 93], [Wand & Jones 95]), daß es bei diesen Schätzern zu hohen Fehlern am Rand kommen kann. Da dieses Phänomen auch bei der Selektivitätsschätzung auftritt - vgl. Kapitel 3.4.2 - wird das Problem der Randbehandlung hier ausführlich behandelt. Der Grund von Randproblemen liegt darin, daß der Kernschätzer auf einer stetigen Schätzung der wahren Dichte beruht, diese aber am Rand abgeschnitten wird. Der Schätzer jedoch setzt die Schätzfunktion stetig über den Rand hinaus fort, was zu einem Verlust an Masse im Randbereich führt. Daraus ergibt sich sowohl, daß das Integral über die (abgeschnittene) Schätzfunktion nicht mehr zum Wert 1 integriert, als auch daß der Schätzer am Rand nicht mehr konsistent ist [Jones 93]. Des weiteren kann gezeigt werden (siehe z.B. [Jones 93]), daß der Bias am Rand stark ansteigt.

In der statistischen Literatur wurden verschiedene Verfahren vorgeschlagen, um diesem Randproblem zu begegnen. Dazu gehören sowohl einfache Methoden, die die Konsistenz des Schätzers auch am Rand wiederherstellen, wie z.B. Renormalisierung oder Spiegelung, als auch Verfahren die mittels spezieller Randkerne sowohl zu einem konsistenten Kernschätzer am Rand führen als auch den Bias am Rand vermindern - für einen Überblick vgl. [Jones 93]. In dieser Arbeit wird die Spiegelung verglichen mit einem von [Dong & Simonoff 94] vorgeschlagenen speziellen Randkern.

Um das Randproblem zufriedenstellend zu lösen, sind möglichst zwei Forderungen zu erfüllen. Erstens muß der Kernschätzer auch am Rand konsistent sein und zweitens sollte der Bias des Schätzers am Rand dem Bias im Innern der Domäne entsprechen. Die erste Forderung kann durch Renormalisierung bzw. durch Spiegelung erzielt werden, wobei der Bias des Kernschätzers am Rand jedoch von der Ordnung her nicht verbessert wird. Diese beiden Verfahren werden im folgenden kurz vorgestellt.

Im folgenden sei der Kern K eine normierte Kernfunktion (siehe Definition 2.16) zweiter Ordnung mit Support $[-1;1]$. Der Einfachheit halber wird zunächst nur der linke Rand l betrachtet mit der Annahme, daß dieser an der Stelle 0 liege. Des weiteren liege x im Randbereich, d.h. es sei $0 <= x < h$ mit $x = qh$, $0 <= q <1$.

Jetzt gilt für den Erwartungswert des Kernschätzers am Randpunkt x:

$$E(\hat{f}(x, h)) = \int_0^\infty \frac{1}{h} K_h(x - y) f(y) dy$$

$$= \frac{1}{h} \int_{x/h}^{\frac{x-\infty}{h}} K(t) f(x-ht) d(x-ht) \text{ (Substitution } y = x-ht)$$

$$= -\int_{x/h}^{\frac{x-\infty}{h}} K(t) f(x-ht) dt = \int_{\frac{x-\infty}{h}}^{x/h} K(t) f(x-ht) dt$$

$$= \int_{-1}^{x/h} K(t) f(x-ht) dt = \int_{-1}^{q} K(t) f(x-ht) dt \text{ (Support von } K\text{; } q = x/h)$$

$$= f(x) \int_{-1}^{q} K(t) dt + -f'(x) ht \int_{-1}^{q} tK(t) dt + f''(x) h^2 t^2 \int_{-1}^{q} t^2 K(t) dt \pm \ldots \qquad (2.64)$$

(Taylor-Entwicklung von f um x)

Für $x \geq h$ wäre $q \geq 1$ und die obere Grenze des Integrals liesse sich wegen des auf $[-1;1]$ beschränkten Supports von K auf 1 setzen. In diesem Fall fällt der zweite Term in obiger Gleichung weg, nicht aber der dritte und der Schätzer konvergiert mit Voraussetzung (2.63) von der Ordnung $O(h^2)$ gegen $f(x)$. Eine solche Aussage läßt sich aber nicht mehr für $q < 1$ treffen, da hier bereits der zweite Term ungleich 0 und der erste Term ungleich $f(x)$ sein kann.

Als nützlich zeigt sich im weiteren Verlauf die folgende abkürzende Schreibweise bei gegebener Kernfunktion K:

$$a_j(K, q) = \int_{-1}^{q} t^j K(t) dt \text{ für } 0 < q < 1. \qquad (2.65)$$

Der gesuchte Rankkern K^l hat also möglichst mehrere Bedingungen zu erfüllen: Er sollte die Konsistenz des Schätzers gewährleisten, einen Bias von $O(h^2)$ bewirken und an der Stelle $x = h$ stetig in die normale Kernfunktion K übergehen. Unter den folgenden Voraussetzungen an den (linken) Randkern (mit Rand $l = 0$) lassen sich diese Ziele erreichen. Dabei sind a) und b) Voraussetzungen, die die Konsistenz des Schätzers gewährleisten, c) und d) hinreichende Voraussetzungen eines für einen Bias von $O(h^2)$ und e) besagt, die der Randkern an der Stelle $x = h$ stetig in die normale Kernfunktion K (2. Ordnung) übergeht.

a) $a_0(K^l, q) = \int_{-1}^{q} K^l(t, q) dt = 1$

b) $\int_{-1}^{q} (K^l(t, q))^2 dt < \infty$

c) $a_1(K^l, q) = \int_{-1}^{q} tK^l(t, q)dt = 0$ (2.66)

d) $a_2(K^l, q) = \int_{-1}^{q} t^2 K^l(t, q)dt \neq 0$

e) $K^l(x, q) \to K(x)$ für $q \to 1$

(Im multivariaten Fall ist $1 = (1, ..., 1)$, $-1 = (-1, \ldots, -1)$, $q = (q_1, ..., q_d)$ und $t = (t_1, ..., t_d)$ definiert.)

Renormalisierung

Eine Möglichkeit, um einen konsistenten Randkernschätzer zu erhalten, besteht darin direkt zu erzwingen, daß die am Rand abgeschnittene Kernfunktionen K auch am Rand zu 1 integriert.

Sei dazu $a_0(K, q)$ wie in Gleichung (2.65) definiert. Dann ist der *renormalisierte Randkern* K_R^l am linken Rand $l = 0$ definiert als

$$K_R^l(x) = K(x) / a_0(K, q) \quad (2.67)$$

Man sieht leicht, daß hiermit obige Voraussetzungen a) und b) erfüllt sind. Da außerdem $a_0(K, q) \to 1$ für $q \to 1$ und $a_0(q) = 1$ für $q \geq 1$ gilt, ist auch obige Voraussetzung e) erfüllt.

Somit ist der *renormalisierte Randkernschätzer* $\hat{f}_R(x)$ definiert als

$$\hat{f}_R(x) = \hat{f}_K(x) / a_0(q) \quad (2.68)$$

Man beachte dabei, daß q von x abhängt. Wie oben gezeigt, handelt es sich bei $\hat{f}_R(x)$ um einen konsistenten Schätzer. Allerdings konvergiert der Bias mit der Ordnung $O(h)$, vergleiche auch [Jones 93]. Ein Bias dieser Ordnung ist jedoch unbefriedigend im Vergleich zum Bias der Ordnung $O(h^2)$ im Inneren, siehe unten.

Spiegelung

Auch Ziel dieses Verfahrens ist es, die fehlende Masse im Randbereich aufzufüllen. Dazu wird durch Spiegelung der Randpunkte und Einbeziehung dieser gespiegelten Randpunkte in die Schätzung realisiert.

Der *gespiegelte Randkern* K_S^l ist am linken Rand $l = 0$ somit definiert als:

$$K_S^l(x) = K(x) + K(-x) \tag{2.69}$$

Dies führt zum *gespiegelten Randkernschätzer* $\hat{f}_S(x)$ mit

$$\hat{f}_S(x) = \hat{f}_K(x) + \hat{f}_K(-x) = \frac{1}{n}\sum_{i=1}^{n}(K_h(x - X_i) + K_h(-x - X_i)) \tag{2.70}$$

Mit K einem Kern 2. Ordnung und einfachen Transformationen läßt sich leicht die Voraussetzung (2.66) a) für den gespiegelten Randkern K_S^l nachprüfen (vgl. auch [Jones 93]):

$$\int_{-1}^{q} K_S{}^l(t, q)dt = \int_{-1}^{q} K(t) + K(-t)dt = \int_{-1}^{q} K(t)dt + \int_{-1}^{q} K(-t)dt$$

$$= \int_{-1}^{q} K(t)dt + \int_{q}^{1} K(t)dt = \int_{-1}^{1} K(t)dt = 1$$

Bedingung b) folgt trivialerweise, so daß die Konsistenz des Schätzers $\hat{f}_S(x)$ am Rand gesichert ist.

In [Jones 93] ist außer der Konsistenz noch gezeigt, daß der Bias von der Konvergenzordnung $O(h)$ ist, was immer noch unbefriedigend ist. Aufgrund der Einfachheit der Berechnung wird dieser Schätzer jedoch zur Selektivitätsschätzung in dieser Arbeit herangezogen und mit einem weiteren konsistenten Randkernschätzer verglichen.

Wegen der Symmetrie der hier betrachteten Kernfunktionen K gilt:

$$K_h(-x - X_i) = K_h(-(x + X_i)) = K_h(x + X_i) = K_h(x - (-X_i)) \tag{2.71}$$

Daher entspricht dieses Verfahren dem Vorgehen, zusätzliche Stichprobenelemente außerhalb des Definitionsbereichs hinzuzufügen, indem die sich im Randbereich befindlichen Stichprobenelemente am Rand gespiegelt werden, vgl. auch z.B. [Silverman 86].

Im univariaten Fall mit linkem Rand l bzw. rechtem Rand r bedeutet dies, daß weitere Stichprobenelemente in die Bereiche $[l-h,l]$ bzw. $[r,r+h]$ eingefügt werden, indem die Stichprobenelemente aus den Bereichen $[l,l+h]$ bzw. $[r-h,r]$ an den Rändern l bzw. r gespiegelt werden.

Abbildung 2.14 zeigt den Einfluß der Spiegelung einer einzelnen Epanechnikow-Kernfunktion an der Stelle $0 < X_i < h$ mit linkem Rand $l = 0$. Die verlorene Masse wird der Kernfunktion wieder hinzugefügt.

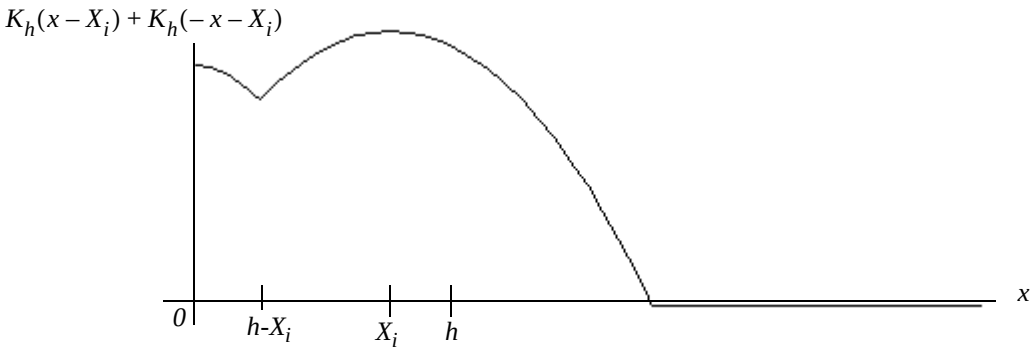


Abbildung 2.14: Überlagerung einer Kernfunktion und ihrer Spiegelung am linken Rand $l = 0$ für $x \geq l$

Der gespiegelte Randkerndichteschätzer $\hat{f}_S(x)$ bei gegebenen Rändern l und r ist nun definiert als

$$\hat{f}_S(x) = \frac{1}{n}\sum_{i=1}^{n}\left(K_h(x-X_i) + K_h(x+X_i-2l) + K_h(x+X_i-2r)\right) \tag{2.72}$$

Abbildung 2.15 zeigt das Verhalten einer Kerndichteschätzung bei einer Stichprobe von 5 Elementen im linken Randbereich von $l = 0$ bis $h = 1$. Verwendet wurde der Epanechnikow-Kern. Die dünne durchgezogene Linie zeigt die Kerndichteschätzung ohne Randbehandlung und die dicke gestrichelte Linie mit Spiegelung. Man sieht deutlich, wie die verlorene Masse am Rand wieder hinzugefügt wird. Für $x \geq h$ sind die beiden Funktionen identisch.

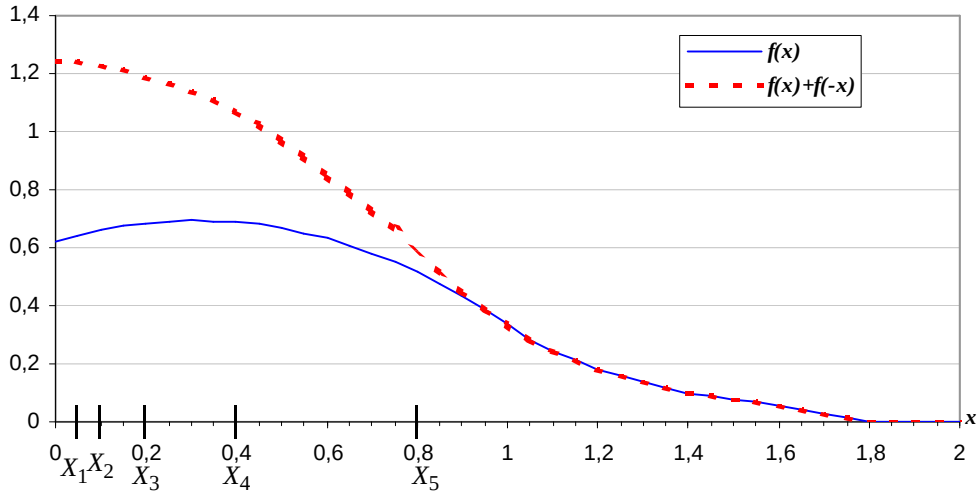


Abbildung 2.15: Beispiel einer Kerndichteschätzung mit Bandbreite $h = 1$ und 5 Stichprobenwerten an den Stellen $X_1 = 0{,}05$, $X_2 = 0{,}10$, $X_3 = 0{,}20$, $X_4 = 0{,}40$ und $X_5 = 0{,}80$

Der gespiegelte Randkernschätzer läßt sich ohne weiteres auf den multivariaten Fall erweitern. Dazu sei zunächst der bivariate Fall betrachtet. Sei dazu der bivariate gespiegelte Randkern $K_S^{\,l}(x) = K_S^{\,l}(x_1, x_2)$ definiert als:

$$K_S^{\,l}(x) = K(x_1, x_2) + K(-x_1, x_2) + K(x_1, -x_2) + K(-x_1, -x_2) \tag{2.73}$$

Dann ergibt sich aus folgender Rechnung, daß Bedingung (2.66) a) erfüllt ist:

$$\int_{-1}^{q} K_S^{\,l}(t)dt = \int_{-1}^{q_2}\int_{-1}^{q_1} K_S^{\,l}(t_1, t_2)dt_1 dt_2$$

$$= \int_{-1}^{q_2}\int_{-1}^{q_1} K(t_1, t_2) + K(-t_1, t_2) + K(t_1, -t_2) + K(-t_1, -t_2)dt_1 dt_2$$

$$= \int_{-1}^{q_2}\int_{-1}^{1} K(t_1, t_2) + K(t_1, -t_2)dt_1 dt_2$$

$$= \int_{-1}^{1}\int_{-1}^{q_2} K(t_1, t_2) + K(t_1, -t_2)dt_2 dt_1 = \int_{-1}^{1}\int_{-1}^{1} K(t_1, t_2)dt_2 dt_1 = \int_{-1}^{1} K(t)dt = 1\,.$$

Da Bedingung b) trivialerweise erfüllt ist, ist somit die Konsistenz des Schätzers am Rand auch im bivariaten Fall gesichert. Aufgrund der Symmetrieeigenschaft des bivariaten Produktkerns K ergibt sich:

$$K_h(-x_1 - X_{i1}, x_2 - X_{i2}) = K_h(x_1 - (-X_{i1}), x_2 - X_{i2}),$$

$$K_h(x_1 - X_{i1}, -x_2 - X_{i2}) = K_h(x_1 - X_{i1}, x_2 - (-X_{i2})) \text{ und}$$

$$K_h(-x_1 - X_{i1}, -x_2 - X_{i2}) = K_h(x_1 - (-X_{i1}), x_2 - (-X_{i2})).$$

Dies entspricht somit der Spiegelung der Stichprobenelemente im Randbereich zuerst an der x_1-Achse und dann - inkl. der bereits gespiegelten - an der x_2-Achse.

Abbildung 2.16 zeigt die Spiegelung eines Stichprobenelementes $X = (X_1, X_2)$ im Teilrandbereich $[(0,0),(h_1,h_2)]$ sowie den Support des zugehörigen Kerns. Die im Intervall $[(X_1 - h_1,0),(0,X_2 + h_2)]$ verlorene Masse wird auf den Bereich $[(0,0),(h_1 - X_1,X_2 + h_2)]$ zurückgespiegelt, die im Intervall $[(0,X_2 - h_2),(X_1 + h_1,0)]$ verlorene Masse auf den Bereich $[(0,0),(X_1 + h_1,h_2 - X_2)]$ und die im Intervall $[(X_1 - h_1,X_2 - h_2),(0,0)]$ verlorene Masse auf den Bereich $[(0,0),(h_1 - X_1,h_2 - X_2)]$.

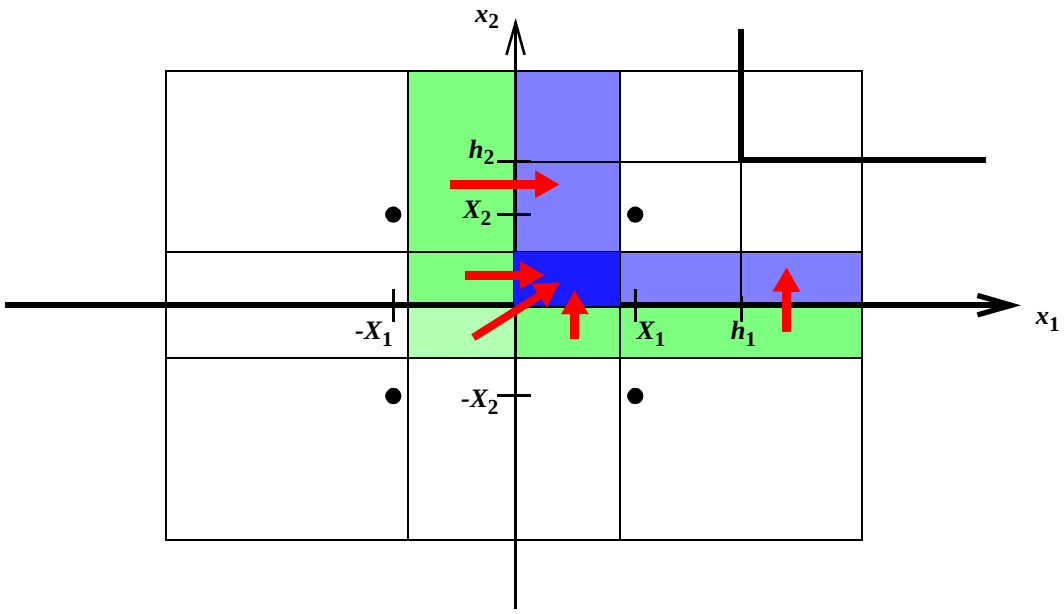


Abbildung 2.16: Spiegelung eines Stichprobenelementes im bivariaten Fall.

Der gespiegelte Randkernschätzer mit linkem Rand l und rechtem Rand r ergibt sich nun mit Anwendung des Produktkernes in folgender Form:

$$\hat{f}_S(x) = \frac{1}{n}\sum_{i=1}^{n} (K_H(x_1 - X_{i1}, x_2 - X_{i2}) + K_H(x_1 + X_{i1} - 2l_1, x_2 + X_i - 2l_2) \quad (2.74)$$

$$+ K_H(x_1 + X_{i1} - 2l_1, x_2 - X_{i2}) + K_H(x_1 - X_{i1}, x_2 + X_{i2} - 2l_2)$$

$$+ K_H(x_1 + X_{i1} - 2r_1, x_2 + X_i - 2r_2) + K_H(x_1 + X_{i1} - 2r_1, x_2 - X_{i2})$$

$$+ K_H(x_1 - X_{i1}, x_2 + X_{i2} - 2r_2) + K_H(x_1 + X_{i1} - 2l_1, x_2 + X_i - 2r_2)$$

$$+ K_H(x_1 + X_{i1} - 2r_1, x_2 + X_i - 2l_2))$$

Die Betrachtungen lassen sich nun auf den multivariaten Fall ausdehnen. Dazu ist der multivariate gespiegelte Randkern definiert als:

$$K_S^{\,l}(x) = \sum K(\pm x_1, \ldots, \pm x_2), \quad (2.75)$$

wobei die Summe über alle möglichen Kombinationen des Vorzeichens geht.

Die weiteren Betrachtungen folgen analog.

Konsistenter Randkernschätzer mit niedrigem Bias

Gesucht ist ein konsistenter Randkernschätzer mit Bias von $O(h^2)$. Der Randkern sei hier mit K_B bezeichnet. Sei dazu $a_l(q) = a_l(K_B^l, q)$ definiert wie in Gleichung (2.65) bei Betrachtung des linken Randes 0. Die Idee ist nun, K_B durch eine Linearkombination der Kernfunktion K und einer dazu in Beziehung stehenden weiteren Funktion L zu bilden, so daß gilt:

$$a_0(K_B^l, q) = 1 \text{ und } a_1(K_B^l, q) = 0. \quad (2.76)$$

Sei weiterhin $c_l(q) = a_l(L, q) = \int_{-1}^{q} x^j L(x)dx$. Dann kann gezeigt werden, daß die Linearkombination

$$K_B^l(x) = \frac{c_1(q)K(x) - a_1(q)L(x)}{c_1(q)a_0(q) - a_1(q)c_0(q)} \quad (2.77)$$

zum gewünschten Bias von $O(h^2)$ führt, vgl. [Jones 93]. Eine einfache Wahl für L ist $L(x) = xK(x)$, vgl. [Jones 93] und [Wand & Jones 95]. Dies führt zu

$$K_B^l(x) = \frac{a_2(q) - a_1(q)x}{a_2(q)a_0(q) - a_1(q)a_1(q)}K(x) \quad (2.78)$$

Andere Randkerne finden sich z.B. in [Gasser & Müller 79]. Es sei darauf hingewiesen, daß diese Randkerne evtl. negative Werte annehmen können. In [Jones & Foster 96] werden spezielle nicht-negative konsistente Randkerne mit Bias $O(h^2)$ vorgestellt.

In [Dong & Simonoff 94] wird ein weiterer einfacher konsistenter Randkern mit Bias $O(h^2)$ vorgeschlagen. Dieser eignet sich aufgrund seiner Einfachheit besonders für praktische Anwendungen und wird in dieser Arbeit in Kapitel 3.4.2 zur Selektivitätsschätzung angewendet. Vorteile sind weiterhin, daß er zum einen auf dem Epanechnikow-Kern beruht und zum anderen einfacher zu integrieren ist als andere Randkerne.

Der Epanechnikow-Randkern $K_{DS}(t,q)$ von [Dong & Simonoff 94] ist am linken bzw. rechten Rand definiert als:

$$K_{DS}^{l}(t,q) = \begin{cases} 3\dfrac{1+q^2-2t^2}{(1+q)^3}, & \text{falls } -1 \le t \le q \\ 0 & \text{sonst} \end{cases} \quad \text{mit } q = \frac{x-l}{h} \text{ und} \tag{2.79}$$

$$K_{DS}^{r}(t,q) = \begin{cases} 3\dfrac{1+q^2-2t^2}{(1+q)^3}, & \text{falls } -q \le t \le 1 \\ 0 & \text{sonst} \end{cases} \quad \text{mit } q = \frac{r-x}{h} \text{ und} \tag{2.80}$$

(vgl. Abbildung 2.17 für verschiedene q)

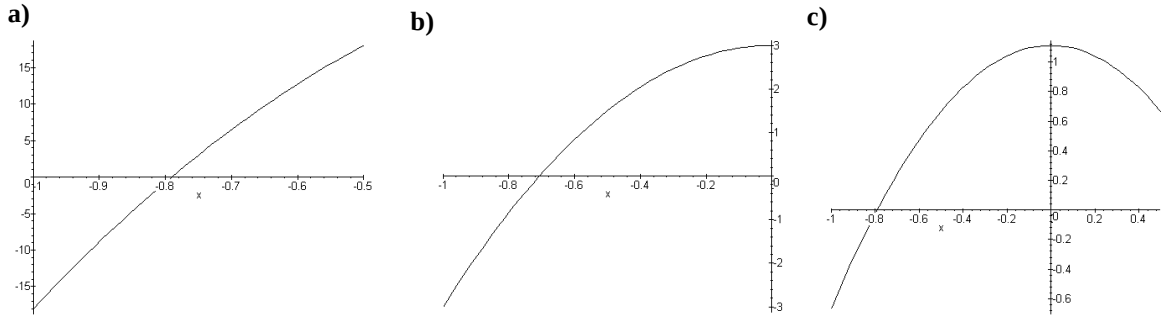


Abbildung 2.17: Epanechnikow-Randkernfunktion von [Dong & Simonoff 94] am linken Rand mit fester Bandbreite $h = 1$ und Support $[-1,q]$ für **a)** $q = -0{,}5$, **b)** $q = 0$ und **c)** $q = 0{,}5$.

Man beachte, daß dieser Randkern für $q < 1$ nicht mehr die Eigenschaften einer Dichtefunktion erfüllt. Für $q = 1$ ergibt sich der normale Epanechnikow-Kern aus Gleichung (2.44).

Für die Dichteschätzung mit linkem Rand l und rechtem Rand r empfiehlt sich eine Fallunterscheidung:

Für $x \in [l + h, r - h]$ berechne $\hat{f}_S(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$.

Für $x \in (l, l + h)$ berechne $\hat{f}_S(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K_{DS}\left(\frac{x - X_i}{h}, q\right)$ mit $q = \frac{x - l}{h}$.

Für $x \in (r - h, r)$ berechne $\hat{f}_S(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K_{DS}\left(\frac{x - X_i}{h}, q\right)$ mit $q = \frac{r - x}{h}$.

Eine andere Möglichkeit besteht darin, den Randkern mit entsprechender Fallunterscheidung zu definieren:

$$K_{RK}(x) = \begin{cases} K^{l}_{DS}(x, q) & \text{für } x \in (l, l + h) \text{ mit } q = \frac{x - l}{h} \\ K(x) & \text{für } x \in [l + h, r - h] \\ K^{r}_{DS}(x, q) & \text{für } x \in (r - h, r) \text{ mit } q = \frac{r - x}{h} \end{cases} \tag{2.81}$$

Diese Definition läßt sich auch für die übrigen Randkerne anwenden.

In den folgenden Abbildungen ist für einen Beispieldatensatz der Unterschied zwischen einer Kerndichteschätzung mit und ohne Randkern dargestellt. Dabei wurde der Beispieldatensatz so gewählt, daß viele Stichprobenwerte am linken Rand $l = 0$ liegen und die Dichte zum rechten Rand hin abnimmt. Die Verteilung entspricht grob einer Exponentialverteilung. Das Beispiel basiert auf einer Stichprobengröße von $n = 20$. Verwendet wurde der Epanechnikow-Kern bzw. der Dong-Simonoff-Randkern. Die Bandbreite wurde in beiden Fällen auf $h = 15$ gesetzt. In den Abbildungen zeigt die fett gezeichnete Linie die jeweilige Dichteschätzung dar, während die dünn gezeichneten Linien die Kernfunktionen an den entsprechenden Stützstellen darstellen. Abbildung 2.18 zeigt die Kernschätzung ohne Randbehandlung. Hier zeigt sich deutlich, daß der Schätzer am linken Rand aufgrund der fehlenden Masse unterschätzt. Wie in Abbildung 2.19 ersichtlich ist, wird dieser Fehler durch Verwendung des Dong-Simonoff-Randkerns vermieden.

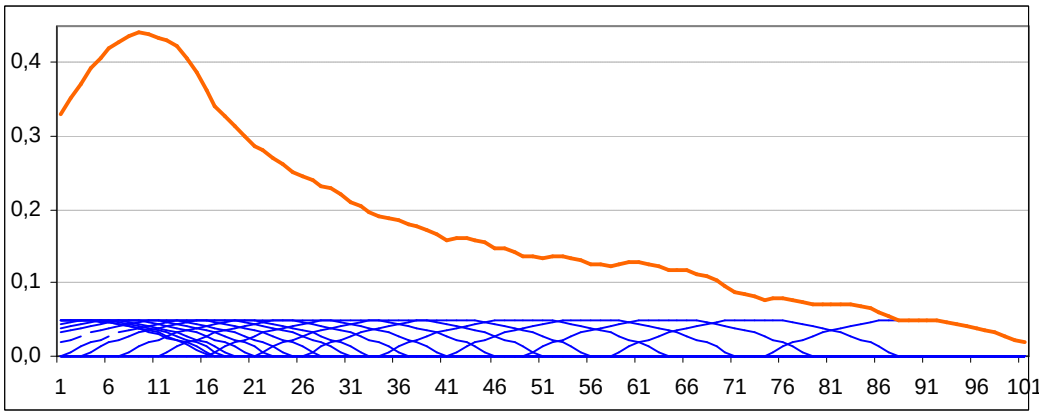


Abbildung 2.18: Dichteschätzung eines Beispieldatensatzes mit Epanechnikow-Kernen ohne Randbehandlung.

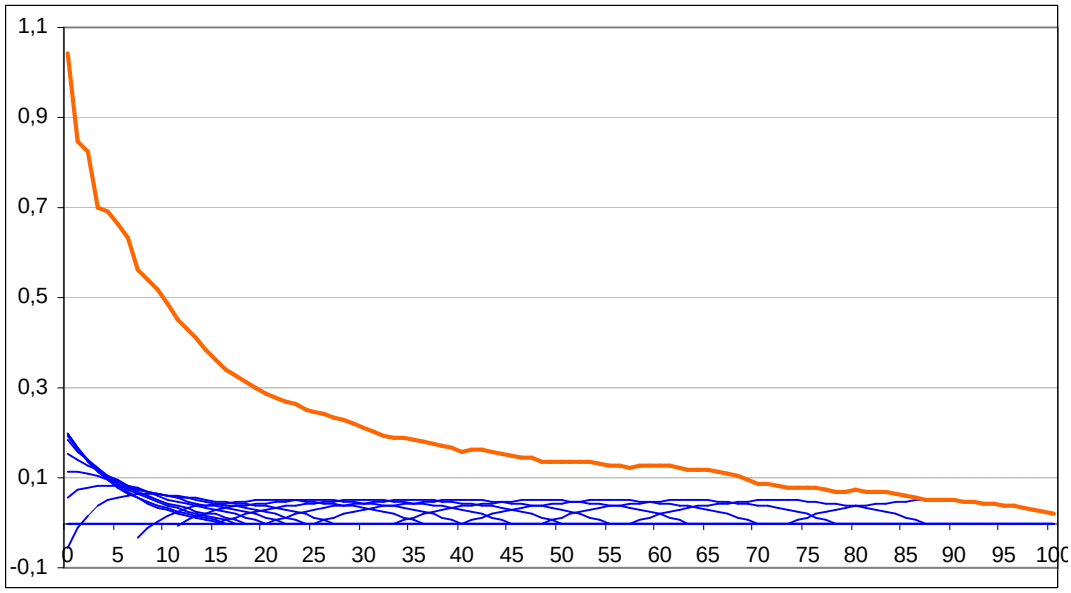


Abbildung 2.19: Dichteschätzung eines Beispieldatensatzes mit Dong-Simonoff-Randkernen.

Auch zu diesen Randkern mit niedrigem Bias lassen sich im multivariaten Fall die entsprechenden Produktkerne mittels Gleichung (2.81) bilden. Z.B. ergibt sich im bivariaten Fall für $x = (x_1,x_2)^t$ mit $x_1 \in (l_1, l_1 + h_1)$ und $x_2 \in [l_2 + h_2, r_2 - h_2]$ der Produktkern $K(x) = K^l_{DS}(x_1, q_1)K(x_2)$ mit $q_1 = (x_1 - l_1) / h_1$, oder für $x_1 \in (r_1, r_1 - h_1)$ und $x_2 \in (l_2, l_2 + h_2)$ der Produktkern $K(x) = K^r_{DS}(x_1, q_1)K^l_{DS}(x_2, q_2)$ mit $q_1 = (r_1 - x_1) / h_1$ und $q_2 = (x_2 - l_2) / h_2$.

Die hier vorgestellten Randkerne (Renormalisierung, Spiegelung, Randkern mit niedrigem Bias) werden in Kapitel 3.4.2 zur Selektivitätsschätzung erweitert.

2.3.6 Weitere nicht-parametrische Schätzverfahren

Es existieren eine Reihe weiterer nicht-parametrischer Schätzverfahren vgl. z.B. [Silverman 86]. Dazu gehören Verfahren unter Verwendung von Polynomapproximation, Spline-Funktionen, Orthogonalreihen oder Maximum penalized likelihood Methoden.

Bei Orthogonalreihenschätzern beispielsweise wird die Dichte geschätzt, indem die Koeffizienten einer Reihenentwicklung geschätzt werden. Dazu läßt sich die Dichtefunktion f umformen zu $\sum_{j=0}^{\infty} f_j \varphi_j$, wobei $f_j = \int_l^r f(x)\varphi_j(x)dx$ für alle i und gewisse Bedingungen an die Koeffizienten φ_i gelten müssen. Häufig wird dazu die Fourierentwicklung genutzt, d.h. die φ_i repräsentieren die Fourierkoeffizienten. Es gibt aber auch Ansätze unter Benutzung anderer Funktionen wie Hermite-Polynome oder mittels Wavelets. Die Schätzung erfolgt unter Ausnutzung einer vorliegenden Stichprobe $X_1,...,X_n$ durch Abschneiden der Restglieder der Reihenentwicklung. Die Anzahl der Terme, nach denen abgeschnitten wird, beeinflußt die Glättung der Schätzung. Der Nachteil des beschriebenen Verfahrens liegt darin, daß möglicherweise negative Werte entstehen können, was der Eigenschaft einer Dichtefunktion widerspricht, vgl. auch [Silverman 86]. Zudem können Oszillationen der Schätzfunktion auftreten.

Die meisten der erwähnten nicht-parametrischen Verfahren lassen sich als Spezialfälle eines allgemeinen Ansatzes unter Verwendung einer generellen Gewichtsfunktion $w(x,X_i)$ darstellen, vgl. z.B. [Silverman 86]. Der allgemeine Schätzer schreibt sich dann als $\hat{f}_G(x) = n^{-1}\sum_{i=1}^{n} w(x, X_i)$. In [Müller 88] wird z.B. gezeigt, daß sich unter gewissen Voraussetzungen Orthogonalreihenschätzer in Kernschätzer überführen lassen.

3. Selektivitätsschätzer

In diesem Kapitel werden die im vorigen Kapitel vorgestellten nicht-parametrischen Dichteschätzer angewendet zur Selektivitätsschätzung in Datenbanksystemen. In allen Fällen wird davon ausgegangen, daß eine Relation R mit einer großen Instanz I von N Tupeln vorliegt. Hiervon wird eine einfache Zufallsstichprobe $X_1,...,X_n$ gezogen mit n Elementen. Es werden sowohl multivariate als auch univariate Selektivitätsschätzer betrachtet.

3.1 Direkte Selektivitätsschätzung mittels Stichprobe

Definition 3.1 (*Direkter Selektivitätsschätzer*):

Sei R eine Relation der Stelligkeit d und I eine Instanz von R sowie $X_1,...,X_n$ eine einfache Zufallsstichprobe der Größe n aus der Instanz. Desweiteren sei eine Anfrage $Q = Q(a,b)$ gegeben. Dann ist der *direkte Selektivitätsschätzer* $\hat{\sigma}_D$ definiert als

$$\hat{\sigma}_D(Q) = \frac{\left|\{a \le X_i \le b\}\right|}{n} \tag{3.1}$$

Dabei ist die Ungleichung $a \le X_i \le b$ für $d > 1$ komponentenweise zu sehen.

Bemerkung 3.1:

Im Falle der Dimension $d = 1$ läßt sich der direkte Selektivitätsschätzer auch leicht mit Hilfe der empirischen Verteilungsfunktion ausdrücken:

$$\hat{\sigma}_{D,1}(Q) = \frac{\left|\{a < X_i \le b\}\right|}{n} = \frac{\left|\{X_i \le b\}\right|}{n} - \frac{\left|\{X_i \le a\}\right|}{n} = F_n(b) - F_n(a) \tag{3.2}$$

Dieses Verfahren entspricht somit einer Selektivitätsschätzung anhand einer Schätzung der empirischen Verteilungsfunktion im Sinne von Gleichung (2.9). Gleichung (3.2) gilt so nicht für d > 1.

3.2 Selektivitätsschätzung mittels Histogrammen

In diesem Abschnitt wird zunächst der Histogrammselektivitätsschätzer motiviert und definiert. Danach werden verschiedene Ausprägungen des Histogrammselektivitätsschätzers zur Selektivitätsschätzung univariater und multivariater Bereichsanfragen vorgestellt. Diese umfassen neben den klassischen Equi-Width- und Equi-Depth-Histogrammschätzern im univariaten Fall den Max-Diff-Histogrammschätzer und im multivariaten Fall die Selektivitätsschätzung mittels Z-Kurve, mittels KD-Baum und mittels Voronoi-Diagrammen.

3.2.1 Allgemeine Selektivitätsschätzung mit Histogrammen

Sei R eine Relation der Stelligkeit d, I eine Instanz aus R und $X_1,...,X_n$ eine einfache Zufallsstichprobe mit n Elementen aus I sowie $Q(a,b)$ eine Anfrage. Um nun Histogrammschätzer zur Selektivitätsschätzung zu verwenden, genügt es Gleichung (2.28) aus der Definition 2.14 des allgemeinen Histogrammdichteschätzers mit den dortigen Voraussetzungen in Gleichung (2.5) einzusetzen:

$$\hat{\sigma}_H(a,b) = \int_a^b \hat{f}_H(t)dt = \frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i}{Vol(C_i)} \cdot \int_a^b I_{C_i}(t)dt. \tag{3.3}$$

Dies führt zu folgender Definition:

Definition 3.2 (*Allgemeiner Histogramm-Selektivitätsschätzer*):

Sei R eine Relation der Stelligkeit d, I eine Instanz von R und $X_1,...,X_n$ eine einfache Zufallsstichprobe der Größe n aus der Instanz. Desweiteren sei eine Anfrage $Q = Q(a,b)$ und eine vollständige, disjunkte Partitionierung der Domäne in k zusammenhängende Bins C_i, i=0..k-1, gegeben. n_i sei für alle i=0..k-1 die Anzahl derjenigen Stichprobenelemente, die in C_i liegen. Der *allgemeine Histogramm-Selektivitätsschätzer* $\hat{\sigma}_H$ ist nun definiert durch

$$\hat{\sigma}_H(a,b) = \int_a^b \hat{f}_H(t)dt = \frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i \cdot \psi_i(Q)}{Vol(C_i)} \text{ mit } \psi_i(Q) = \int_a^b I_{C_i}(t)dt \tag{3.4}$$

und $Vol(C_i)$ wie in Definition 2.14.

Entsprechend den in Kapitel 2.3.2 definierten Varianten des allgemeinen Histogramm-Dichteschätzers (vgl. Tabelle 2.1) ergeben sich hier die folgenden Varianten:

- allg., d-dim. Intervalle $C_i = (c_{i,min}, c_{i,max}]$, $c_i \in \Re^d$:

$$\hat{\sigma}_{H,I}(a,b) = \int_a^b \hat{f}_{H,I}(t)dt = \frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i \cdot \psi_i(Q)}{\prod_{j=1}^{d} h_{ij}} \text{ mit } h_{ij} = c_{i,max,j} - c_{i,min,j} \tag{3.5}$$

- allg., 1-dim. Intervall $C_i = (c_{i,min}, c_{i,max}]$, $c_i \in \Re$:

$$\hat{\sigma}_{H,1}(a,b) = \int_a^b \hat{f}_{H,1}(t)dt = \frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i \cdot \psi_i(Q)}{h_i} \text{ mit } h_i = c_{i,max} - c_{i,min} \tag{3.6}$$

- Equi-Width, allg.:

$$\hat{\sigma}_{EW}(a,b) = \int_a^b \hat{f}_{EW}(t)dt = \frac{1}{n \cdot Vol(C)} \cdot \sum_{i=0}^{k-1} n_i \cdot \psi_i(Q) \text{ mit } Vol(C_i) = const \text{ für alle } i \quad (3.7)$$

- Equi-Width, *d*-dim. Intervalle $C_i = (c_{i,min}, c_{i,max}]$, $c_i \in \Re^d$:

$$\hat{\sigma}_{EW}(a,b) = \int_a^b \hat{f}_{EW}(t)dt = \frac{1}{n \cdot \prod_{j=1}^{d} h_j} \cdot \sum_{i=0}^{k-1} n_i \cdot \psi_i(Q) \quad (3.8)$$

- Equi-Width, 1-dim. Intervall $C_i = (c_{i,min}, c_{i,max}]$, $c_i \in \Re$:

$$\hat{\sigma}_{EW}(a,b) = \int_a^b \hat{f}_{EW}(t)dt = \frac{1}{n \cdot h} \cdot \sum_{i=0}^{k-1} n_i \cdot \psi_i(Q) \quad (3.9)$$

- Equi-Depth, allg.:

$$\hat{\sigma}_{ED}(a,b) = \int_a^b \hat{f}_{ED}(t)dt = \frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{\psi_i(Q)}{Vol(C_i)} \quad (3.10)$$

- Equi-Depth, *d*-dim. Intervalle $C_i = (c_{i,min}, c_{i,max}]$, $c_i \in \Re^d$:

$$\hat{\sigma}_{ED,I}(a,b) = \int_a^b \hat{f}_{ED,I}(t)dt = \frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{\psi_i(Q)}{\prod_{j=1}^{d} h_{ij}} \quad (3.11)$$

- Equi-Depth, 1-dim. Intervall $C_i = (c_{i,min}, c_{i,max}]$, $c_i \in \Re$:

$$\hat{\sigma}_{ED,1}(a,b) = \int_a^b \hat{f}_{ED,1}(t)dt = \frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{\psi_i(Q)}{h_i} \quad (3.12)$$

Auch die verschiedenen Histogrammselektivitätsschätzer seien noch einmal analog zu Tabelle 2.1 der Übersicht halber tabellarisch zusammengefaßt:

	bel. Bins	***d*-dim. Intervalle**	**dim = 1**
allg.	$\frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i \cdot \psi_i(Q)}{Vol(C_i)}$	$\frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i \cdot \psi_i(Q)}{\prod_{j=1}^{d} h_{ij}}$	$\frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{n_i \cdot \psi_i(Q)}{h_i}$
EW	$\frac{1}{n \cdot Vol(C)} \cdot \sum_{i=0}^{k-1} n_i \cdot \psi_i(Q)$	$\frac{1}{n \cdot \prod_{j=1}^{d} h_j} \cdot \sum_{i=0}^{k-1} n_i \cdot \psi_i(Q)$	$\frac{1}{n \cdot h_i} \cdot \sum_{i=0}^{k-1} n_i \cdot \psi_i(Q)$
ED	$\frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{\psi_i(Q)}{Vol(C_i)}$	$\frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{\psi_i(Q)}{\prod_{j=1}^{d} h_{ij}}$	$\frac{1}{k} \cdot \sum_{i=0}^{k-1} \frac{\psi_i(Q)}{h_i}$

Tabelle 3.1: Übersicht über verschiedene Histogramm-Selektivitätsschätzer

Man beachte, daß stets $0 \le \psi_i(Q) \le Vol(C_i)$ gilt mit $\psi_i(Q) = Vol(C_i \cap Q)$. Daraus folgt $\psi_i(Q) = 0$ für $Q \cap C_i = \varnothing$ und $\psi_i(Q) = Vol(C_i)$ für $C_i \subseteq Q$. Mit anderen Worten, liegt ein Bin C_i außerhalb eines Anfragebereichs Q, so liefert es keinen Beitrag zum Selektivitätsschätzer, liegt C_i ganz innerhalb von Q so liefert es vollen Beitrag n_i / n, ansonsten entsprechend des Anteils der Überlappung an Q.

Die Nachteile der Histogramm-Dichteschätzer übertragen sich auch auf die Histogramm-Selektivitätsschätzer, insbesondere ist der Schätzer nach wie vor nicht erwartungstreu und abhängig von der Wahl des Startpunktes.

Die verschiedenen Histogrammselektivitätsschätzer unterscheiden sich im Wesentlichen durch die Art der Partitionierung. Dazu gehören im univariaten Fall der Equi-Width-, der Equi-Depth- und der Max-Diff-Histogrammselektivitätsschätzer, die in Kapitel 5.5 ausführlich anhand von Experimenten untersucht werden. Im multivariaten Fall zählen dazu zum einen der klassische Equi-Width-Histogrammselektivitätsschätzer und zum anderen der Equi-Depth-Histogrammselektivitätsschätzer in der Variante von [Muralikrishna & DeWitt 88], vgl. Kapitel 1.3.3. Diese werden in Kapitel 5.6 ausführlich mittels Experimenten auf bivariaten Testdaten untersucht.

In den nächsten Unterkapiteln werden weitere neue Methoden zur multivariaten Selektivitätsschätzung mit Histogrammen vorgestellt. Bei der multivariaten Selektivitätsschätzung mittels raumfüllender Kurve (Kap. 3.2.2) handelt es sich um die Verwendung univariater Histogrammselektivitätsschätzer auf eindimensionalen Daten, die durch Abbildung mehrdimensionaler Daten auf eine ein-dimensionale Ordnung entstanden sind. Dabei läßt sich die ganze Palette univariater Histogrammschätzer vom Equi-Width-Histogramm über das Equi-Depth-Histogramm bis zum Max-Diff-Histogramm einsetzen. Bei der multivariaten Selektivitätsschätzung mittels KD-Baum (Kap. 3.2.3) wird eine im Bereich der Datenbanksysteme bekannte und verbreitete Indexstruktur - der KD-Baum - verwendet, um eine vollständige, disjunkte Partitionierung zur

Bildung der Histogramm-Bins zu erhalten. Hierauf wird dann der allgemeine multivariate Histogrammselektivitätsschätzer angewendet. Bei der multivariaten Selektivitätsschätzung mittels Voronoi-Diagrammen (Kap. 3.2.4) erhält man mittels der Voronoi-Regionen eine Partitionierung, die zwar vollständig aber nicht disjunkt ist, da sich die Voronoi-Regionen an ihren gemeinsamen Rändern überlappen. Nichtsdestotrotz lassen sich auf diese Weise Klassen bilden, die durch Anwendung des allgemeinen multivariaten Histogrammselektivitätsschätzers zur Selektivitätsschätzung verwendet werden können. Die Selektivitätsschätzung mittels Average Shifted Histogrammen wird in Kap. 3.3 vorgestellt. Auch diese speziellen Histogrammselektivitätsschätzer werden in Kapitel 5.6 ausführlich mittels Experimenten auf bivariaten Testdaten untersucht.

3.2.2 Multivariate Selektivitätsschätzung mittels raumfüllender Kurve

Raumfüllende Kurven wurden zuerst von Peano eingeführt. Aus diesem Grunde werden raumfüllende Kurven oft auch Peano-Kurven genannt. Seit dieser Zeit haben sich zahlreiche Varianten gebildet. Ein Überblick über raumfüllende Kurven findet sich in dem Buch von Sagan [Sagan 94]. Für praktische Zwecke zeichnen sie sich dadurch aus, daß sie es ermöglichen Punkte eines diskreten und beschränkten mehr-dimensionalen Raumes zu ordnen. Ein typisches und in der Datenbank-Literatur häufiger verwendetes Beispiel ist die Lesbesgue-Kurve, wie sie für einen zwei-dimensionalen Raum in Abbildung 3.1 dargestellt ist. Aufgrund ihres Aussehens wird diese Kurve auch oft Z-Kurve genannt.

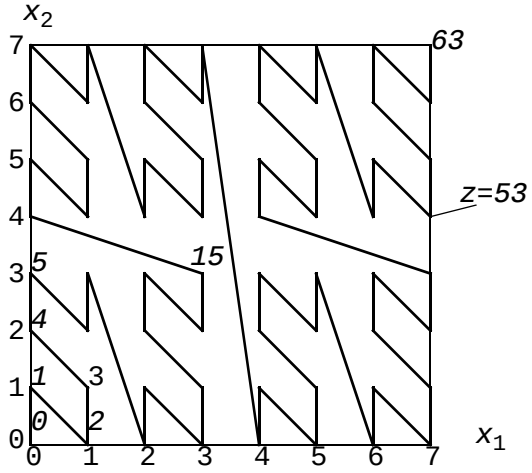


Abbildung 3.1: Z-Ordnung im zweidimensionalen Fall ($p_1 = p_2 = 3$, $|X| = 64$, Z-Werte kursiv)

Im folgenden wird der Einfachheit halber wie im Beispiel davon ausgegangen, daß der zugrunde liegende Datenraum durch einen achsenparallelen Hyperwürfel beschränkt ist und die in ihm enthaltenen diskreten Punkte äquidistant in jeder Dimension sind. Des weiteren wird im folgenden hauptsächlich die Z-Kurve betrachtet, da sie einfache Berechnungsmethoden anbietet um Punkte des Ursprungsraumes auf die entsprechenden Z-Werte und vice versa abzubilden ([Buskamp 97]). Andere raumfüllende Kurven wie die Hilbert-Kurve könnten ebenfalls behan-

delt werden, sind aber komplizierter zu berechnen. Dabei liegt die Schwierigkeit weniger in der Berechnung der entsprechenden Hilbert-Werte als in der Berechnung des Durchschnitts vom Anfragebereich und den ein-dimensionalen Histogrammbins auf der raumfüllenden Kurve. Dies wird weiter unten deutlich. Zunächst sei der Begriff der Z-Ordnung definiert.

Definition 3.3 (*Z-Ordnung*):

Sei $U = U_1 \times ... \times U_d$ gegeben mit $U = \{0 \ldots 2^p - 1\}$, $|U| = 2^{dp}$. Die Komponenten der Elemente $v = (v_1, \ldots, v_d) \in U$ lassen sich darstellen als $v_i = \sum_{j=0}^{p-1} 2^j \cdot a_{ij}$, $a_{ij} \in \{0,1\}$. Dann sei die *Z-Ordnung* von U definiert als bijektive Abbildung ζ von v auf den entsprechenden *Z-Wert* $z \in Z = \{0, \ldots, 2^{dp}-1\}$ in folgender Weise:

$$\zeta : U \rightarrow Z,\ v \rightarrow \zeta(v) := z := \sum_{i=1}^{d} \sum_{j=0}^{p-1} 2^{d \cdot j + i - 1}. \tag{3.13}$$

Bemerkung 3.2:

- Sei ein Z-Wert z wie folgt gegeben: $z = \sum_{j=0}^{p-1} 2^j \cdot c_j$. Dann können die Komponenten x_i des entsprechenden Punktes im Ursprungsraums berechnet werden nach der Gleichung $x_i = \sum_{j=0}^{p-1} 2^j \cdot c_{d \cdot j + i - 1}$, $i = 1..d$.
- In binärer Notation läßt sich Gleichung (3.13) schrieben als $z = (a_{np},...,a_{n0},...,a_{1p},...,a_{10})_2$ mit $x_i = (a_{ip},...,a_{i0})_2$.

Beispiel 3.1:

- Sei $d = 2$, $p = 3$, $x = 7 = 0111_2$ und $y = 4 = 0100_2$, dann ist $z = 00110101_2 = 53$ (vgl. Abb. 3.1).
- Sei umgekehrt $d = 3$, $p = 3$ und $z = 273 = 100010001_2$, dann ist $x_1 = 001_2 = 1$, $x_2 = 010_2 = 2$ und $x_3 = 100_2 = 3$.

Sei nun eine Relation R mit mehr-dimensionalen Attributen sowie eine Instanz I von R mit N Tupeln gegeben. OBdA seien die Werte der Instanz ganzzahlig diskret. (Dies läßt sich aufgrund der endlichen Größe der Instanz immer herstellen.) Die mehrdimensionalen Attribute lassen sich nun mittels der Z-Ordnung auf eine 1-dimensionale Ordnung abbilden. Auf dieser Ordnung lassen sich die bekannten eindimensionalen Histogrammschätzer anwenden. Im folgenden werden im wesentlichen der Equi-Width-, der Equi-Depth- und der Max-Diff-Histogrammschätzer

betrachtet. Beim Max-Diff-Histogrammschätzer kann hier weiter unterschieden werden, ob die Differenzen zwischen den Z-Werten oder zwischen den Ursprungswerten der Tupel berechnet werden.

Man bemerke, daß durch die so auf der Z-Ordnung definierten Histogrammbins i.a. keine Histogrammbins in der Domäne der Relation definiert sind. Nichtsdestotrotz können die auf der Z-Ordnung definierten Histogramme zur Selektivitätsschätzung von Bereichsanfragen in mehrdimensionalen Räumen verwendet werden. Sei dazu eine *d*-dimensionale Anfrage $Q(a,b) = [a,b]$, $a, b \in X$, gegeben, so kann jeder Punkt aus Q auf einen Z-Wert anhand von Gleichung (3.13) abgebildet werden. Im Unterschied zu Q sind die durch die Abbildung $\zeta(Q)$ gebildeten Kurvenstücke nicht notwendig zusammenhängend. Dies ist in Abbildung 3.2 ersichtlich, wo eine Anfragerechteck Q mit den Punkten (2,3), (3,3), (4,3), (2,4), (3,4), (4,4), (2,5), (3,5), (4,5) auf einem Datenbereich von (0,0) bis (7,7) gegeben ist. Die Punkte in Q werden abgebildet mittels (3.13) auf die Z-Werte (nach Umsortierung) 13, 15, 24 bis 27, 37, 48 und 49. Dementsprechend besteht $\zeta(Q)$ aus r, $0 \le r \le |Q|$, jeweils zusammenhängenden Teilstücken. Im Beispiel von Abbildung 3.2 ist $r = 5$ und die Q_i sind $[13,13]$, $[15,15]$, $[24,27]$, $[37,37]$, $[48,49]$.

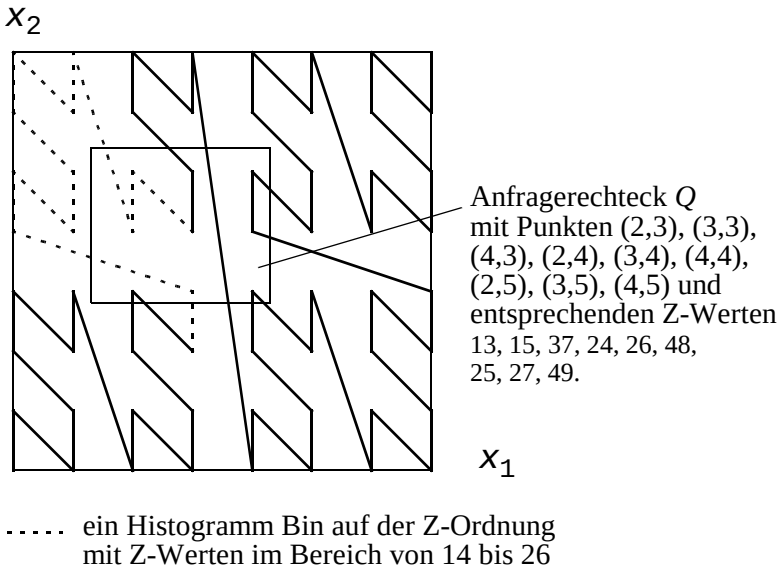


Abbildung 3.2: Anfragefenster Q und Histogramm Bins auf der Z-Kurve

Die Q_i können nun jeweils als separate Anfragen $Q_i \subseteq Z$, $i = 1..r$, behandelt werden, so daß die Selektivitätsschätzung von Q am Ende aus der Selektivitätsschätzung der Q_i aggregiert werden kann:

$$\hat{\sigma}(Q) = \sum_{i=1}^{r} \hat{\sigma}(Q_i). \tag{3.14}$$

Die Berechnung der Q_i ist sehr auf aufwendig. Eine Verbesserung der Laufzeit läßt sich durch einen rekursiven Algorithmus abgeleitet von [Tropf & Herzog 81] erzielen. Dabei wird zunächst, ausgehend von einem Histogrammbin, ein zusammenhängender Bereich der Z-Kurve betrachtet. Anhand gewisser Eigenschaften läßt sich entscheiden, ob der Bereich innerhalb oder außerhalb des betrachteten Anfragebereichs oder diesen schneidet. Im letzten Fall wird der Bereich weiter aufgeteilt, bis zuletzt die Teilstücke identifiziert sind, die komplett im Anfragebereich liegen. Für eine genaue Beschreibung siehe [Tropf & Herzog 81] oder auch [Buskamp 97].

Es lassen sich nun auf der 1-dimensionalen Z-Ordnung verschiedene Typen von univariaten Histogramm-Selektivitätsschätzern (vgl. oben) anwenden. In den Experimenten (Kapitel 5.6) wurden die folgenden Typen mit entsprechender Begründung gewählt:

- Equi-Width- (EW) und Equi-Depth- (ED) Histogramme sind die klassischen Typen, mit denen die Ergebnisse verglichen werden.
- Das Max-Diff-Histogramm kann hier nach zwei Varianten unterschieden werden. Entweder werden die Differenzen bzgl. der Z-Werte (MZ) oder bzgl. der Domänenpunkte *v* (MD) gebildet. Durch die letzte Variante werden innerhalb eines Bins große Bereiche auf der Z-Kurve ohne Stichprobenelemente vermieden, so daß benachbarte Punkte vorzugsweise einem Bin zugeordnet werden.

Algorithmus 3.1 (*multivariate Selektivitätsschätzung mittels Z-Ordnung*):

A. Bilden des Histogrammes:

Gegeben: Relation R der Stelligkeit d mit Instanz I und einfache Zufallsstichprobe $X_1,...,X_n$ der Größe n, die Tupel der Instanz seien oBdA ganzzahlig im Bereich $0,...,2^p - 1$.

1. Berechne für jeden Stichprobenwert X_k den entsprechenden Z-Wert z_k mit (3.13), $k = 1...n$.
2. Bilde anhand der z_k ein univariates Histogramm von einem der obigen Typen und speichere für jedes Histogramm-Bin C_i, $i = 0...k\text{-}1$, die Anzahl n_i der darin enthaltenen z_j.

B. Bearbeiten der Anfrage:

Gegeben: wie A. mit entsprechendem Histogramm (C_i, n_i) sowie eine Bereichsanfrage $Q(a,b)$.

1. Bestimme die r zusammenhängenden Teilanfragen $Q_j = Q_j(a_j,b_j)$ auf der Z-Ordnung.
2. Für alle Histogramm-Bins C_i, $i = 0...k\text{-}1$, und alle Teilanfragen Q_j, $j = 1...r$, auf der Z-Ordnung berechne $\psi_i(Q_j) = \int_{a_j}^{b_j} I_i(t)dt$ und anschließend $\tilde{\sigma}(Q_j) = \frac{1}{n} \cdot \sum_{i=0}^{k-1} n_i \cdot \psi_i(Q_j)/h_i$ mit $h_i = c_{i,max} - c_{i,min} + 1$.

3. Nun ist der Selektivitätsschätzer $\hat{\sigma}$ der Anfrage Q gegeben durch $\hat{\sigma}(Q) = \sum_{j=1}^{r} \tilde{\sigma}(Q_j)$

Hierzu wurden Experimente mit $d = 2$ und verschiedenen Datensätzen sowie unterschiedlichen p gemacht. Die Ergebnisse werden in Kapitel 5.6 vorgestellt.

3.2.3 Multivariate Selektivitätsschätzung mittels KD-Baum-Partitionierung

Der (adaptive) KD-Baum [Robinson 81] ist wohlbekannt aus der Indextheorie mehrdimensionaler Datenbanken. Es handelt sich um einen binären, nicht-balancierten Baum, welcher den Datenraum in nicht-überlappende, achsenparallele (Hyper-)Rechtecke partitioniert. Die Partitionierung erfolgt dabei in rekursiver Form, indem die Rechtecke orthogonal zu einer Achse gesplittet werden, so daß sich eine Hierarchie von Rechtecken ergibt. Die mehrdimensionalen Objekte werden anschließend den Blattknoten zugeordnet[1], während die inneren Knoten die Informationen enthalten über die Split-Achse und den entsprechenden Splitwert als Diskriminator für eine spätere Suche. Bei jedem Split muß entschieden werden, welches Rechteck gesplittet werden soll, orthogonal zu welcher Achse und bei welchem Wert. Nachteil des KD-Baums gegenüber anderen mehrdimensionalen Indexstrukturen wie z.B. dem R-Baum [Guttmann 84] oder seiner Variante R*-Baum ([Beckmann et al. 90], [Seeger 96]) ist, daß die Qualität (bzgl. der Suche) stark abhängt von der gewählten Split-Strategie und der Ordnung der Elemente während des Einfügens in den Baum.

In dieser Arbeit wird die Eigenschaft des KD-Baums genutzt, daß er eine vollständige und disjunkte Partitionierung der Domäne auf allen Ebenen liefert, was bei anderen mehrdimensionalen Indexstrukturen wie dem R-Baum oder R*-Baum nicht gegeben ist. Auf diese Partitionierung kann nun der allgemeine Histogrammselektivitätsschätzer angewendet werden. Dabei werden die in den Blattknoten gespeicherten Rechtecke als Histogramm-Bins verwendet. Ist eine Bereichsanfrage gegeben, so kann der KD-Baum rekursiv durchsucht werden, um die Schnittmenge des Anfragebereichs und der Histogramm-Bins in den Blattknoten zu bestimmen. Anstatt den Blattknoten nun die Objekte selbst zuzuordnen, wird in ihnen lediglich die Anzahl der Objekte in dem zum Blattknoten gehörenden Rechteck gespeichert.

Ein offenes Problem ist die Wahl der Splitstrategie. Sie kann so gewählt werden, daß sich verschiedene Histogramm-Typen ergeben entsprechend den weiter oben vorgestellten. Z.B. werden bei einer Equi-Width bzw. Equi-Depth Splitstrategie die Splitwerte so gewählt, daß die sich ergebenden Rechtecke einer Hierarchiestufe alle das gleiche Volumen haben bzw. die gleiche Anzahl von Elementen besitzen. Hier ist eine andere heuristische Splitstrategie gewählt worden. Die Splitparameter sind so gewählt worden, daß die Varianz (mit euklidischer Norm) minimiert wird. D.h. berechnet man die Varianz der Punkte innerhalb jeder Partition, so wird diejenige

1. Der ursprüngliche KD-Baum unterscheidet sich hiervon etwas, indem dort die Datenpunkte auch in den inneren Knoten gespeichert werden [Samet 90], die hier beschriebene Variante hat sich aber in der Praxis durchgesetzt.

Partition gesplittet, die die größte Varianz besitzt. Der Split erfolgt in der Dimension, in der die größte Varianz in der Randverteilung vorliegt. Geteilt wird in dieser Dimension analog dem MaxDiff-Verfahren an der Mitte zweier benachbarter Punkte, deren Abstand in dieser Dimension von allen benachbarten Punkten am größten ist. Die Hoffnung ist, daß bei einer multimodalen Verteilung in dieser Partition diese in zwei Partitionen mit weniger Modi geteilt wird bis schließlich nur noch unimodale Verteilungen in den einzelnen Partitionen vorliegen. Hierbei ist jedoch das Problem der Nichtseparierbarkeit von manchen Datenverteilungen zu beachten. Der rekursive Algorithmus bricht ab, sobald eine vorgegebene Anzahl von Histogrammbins erreicht. Kriterien für eine optimale Anzahl von Histogrammbins sind nicht bekannt.

Definition 3.4 (*KD-Baum Histogrammselektivitätsschätzer*):

Seien die Voraussetzungen wie in Definition 3.2 erfüllt. Weiterhin sei eine Partitionierung der Domäne durch einen KD-Baum wie oben beschrieben gegeben. Dann ist der *KD-Baum Histogrammselektivitätsschätzer* definiert als der allgemeine Histogrammselektivitätsschätzer bzgl. der KD-Baum Partitionierung.

Algorithmus 3.2 (*KD-Baum Histogrammselektivitätsschätzer*):

A. Bilden des Histogrammes:

Gegeben: Dimension d, eine Relation R mit Instanz I und eine einfache Zufallsstichprobe $X_1,...,X_n$ der Größe n.

1. Erzeuge den KD-Baum mittels einer geeigneten Splitstrategie auf der Stichprobe. Die entstehenden k Blattknoten partitionieren den Datenraum in Histogramm-Bins C_i, $i = 0...k\text{-}1$, wobei k durch die Splitstrategie gegeben ist.
2. Für jedes Histogramm-Bin C_i, $i = 0...k\text{-}1$, in den Blattknoten speichere die Anzahl n_i der Stichprobenwerte in dem zugehörigen Bereich.

B. Bearbeiten der Anfrage:

Gegeben: wie A. mit entsprechendem Histogramm (C_i, n_i) sowie eine Bereichsanfrage $Q(a,b)$.

1. Die Selektivitätsschätzung ergibt sich nun durch Anwendung des multivariaten Intervall-Histogramm-Selektivitätsschätzers wie in Gleichung (3.5) definiert.

Die Ergebnisse des KD-Baum Histogrammselektivitätsschätzers werden in Kapitel 5.6 vorgestellt und mit den anderen Histogramm-Selektivitätsschätzern verglichen.

In [Poosala & Ioannidis 97] führen die Autoren die sogenannte MHIST-p Methode ein. Dabei wird rekursiv jedes Rechteck in $p \geq 2$ weitere Rechtecke partitioniert. Für $p > 2$ ist der resultierende Baum nicht mehr binär. Die Autoren berichten, daß sie die besten Ergebnisse in ihren Experimenten für $p = 2$ erzielt haben, so daß der Fall $p > 2$ nicht weiter berücksichtigt wird. Für $p = 2$ ergibt sich eine Partitionierung analog zum oben vorgestellten KD-Baum. In der Splitstra-

tegie von [Poosala & Ioannidis 97] werden nur Randverteilungen betrachtet. Es wird diejenige Partition geteilt, in der es ein - univariates - Attribut gibt, das bzgl. irgendeinem vorgegebenen Kriterium eine maximale Eigenschaft hat. In der entsprechenden Dimension wird dann geteilt. Als Kriterien werden die in [Poosala et al. 96] vorgestellten Kriterien für univariate Histogramme benutzt. Dagegen gibt es in dem oben vorgestellten KD-Baum Verfahren keine Einschränkungen bzgl. der Splitstrategie, so daß auch multivariate Kenngrößen betrachtet werden können.

3.2.4 Multivariate Selektivitätsschätzung mittels Voronoi-Partitionierung

Voronoi-Diagramme sind aus Graphentheorie bekannt und haben sich in verschiedenen Anwendungen wie z.B. der k-nächsten Nachbar-Suche als sehr nützlich erwiesen [Ottmann & Widmayer 96]. In dieser Arbeit werden sie genutzt, um mit den dabei entstehenden Voronoi-Regionen den zugrundeliegenden Datenraum zu partitionieren.

Sei eine (kleine) Punktmenge $A = \{a_1,...,a_k\} \in \Re^d$ gegeben sowie eine beliebige Metrik D. Ziel eines Voronoi-Diagramms ist es zunächst, eine Partitionierung der Domäne derart zu bekommen, daß jedes Element einer Voronoi-Region bzgl. a_i näher bzgl. D an a_i ist als zu irgendeinem anderen Punkt aus A. Dies ist genauer in folgender Definition ausgedrückt:

Definition 3.5 (*Voronoi-Region*):

Sei eine Metrik D gegeben, ein d-dimensionaler Raum $U \subset \Re^n$ sowie eine endliche Menge $A = \{a_1, ..., a_k\}$ von Punkten aus der Domäne U. Die Punkte a_i heißen *Ankerpunkte.* Dann ist eine *Voronoi-Region* $V(a_i)$ bzgl. eines Ankerpunktes a_i (bzgl. der Metrik D) definiert als

$$V(a_i) = \{v \in U | D(v, a_i) \leq D(v, a_j) \forall a_j \in U, j \neq i\} \tag{3.15}$$

Bemerkung 3.3:

- Jede Voronoi-Region $V(a_i)$ ist zusammenhängend und konvex.
- Sei V die Vereinigung aller Voronoi-Regionen $V(a_i)$, $i = 1..k$. Dann liefert V eine vollständige Partitionierung des Datenraums U: $\bigcup_{i=1}^{k} V(a_i) = U$.
- Die Voronoi-Regionen $V(a_i)$ sind nicht disjunkt, da es Punkte gibt, die zu zwei verschiedenen Ankerpunkte a_i und a_j benachbarter Voronoi-Regionen $V(a_i)$ und $V(a_j)$ gleichen kürzesten Abstand haben. Diese liegen dann sowohl in als $V(a_i)$ auch in $V(a_j)$.
- Die Voronoi-Regionen werden durch eine Menge von Kanten, den *Voronoi-Kanten,* getrennt. Alle Kanten aller Voronoi-Regionen bilden das *Voronoi-Diagramm* (s.u.). Die zugehörigen Knoten heißen *Voronoi-Knoten.*

Definition 3.6 (*Voronoi-Diagramm*):

Sei $E(V(a_i))$ die Menge von Kanten, die zu einer Voronoi-Region $V(a_i)$ gehören, und $N(V(a_i))$ die entsprechende Menge von Knoten bzgl. $V(a_i)$. Das *Voronoi-Diagramm* von U bzgl. der Metrik D ist definiert als der Graph $G=(N,E)$ mit $N = \bigcup_{i=1}^{k} N(V(a_i))$ und $E = \bigcup_{i=1}^{k} E(V(a_i))$.

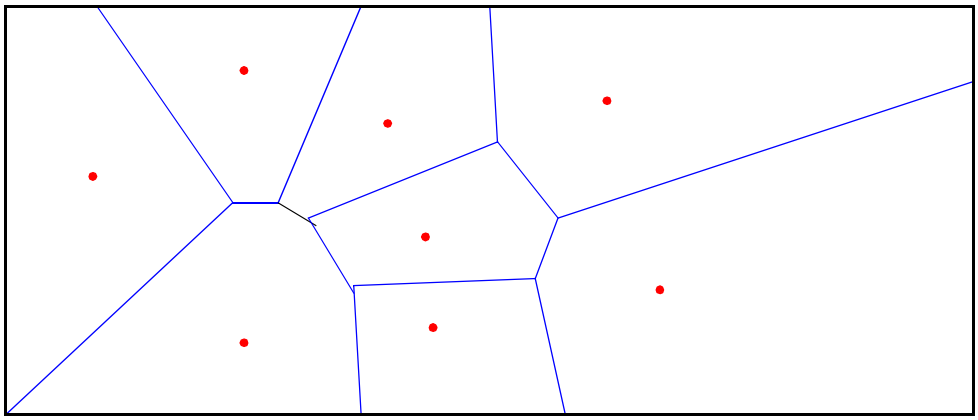


Abbildung 3.3: Voronoi-Diagramm mit 8 Ankerpunkten.

Bemerkung 3.4:

Das Voronoi-Diagramm steht in direktem Bezug zur *Delaunay-Triangulation*: Unter der Annahme, daß zu jedem Ankerpunkt maximal drei Voronoi-Regionen gehören, erhält man die Delaunay-Triangulation aus dem Voronoi-Diagramm, indem man diejenigen Ankerpunkte miteinander verbindet, die eine gemeinsame Voronoi-Kante besitzen. Jedes Voronoi-Diagramm kann in eine Delaunay-Triangulation transformiert werden und umgekehrt.

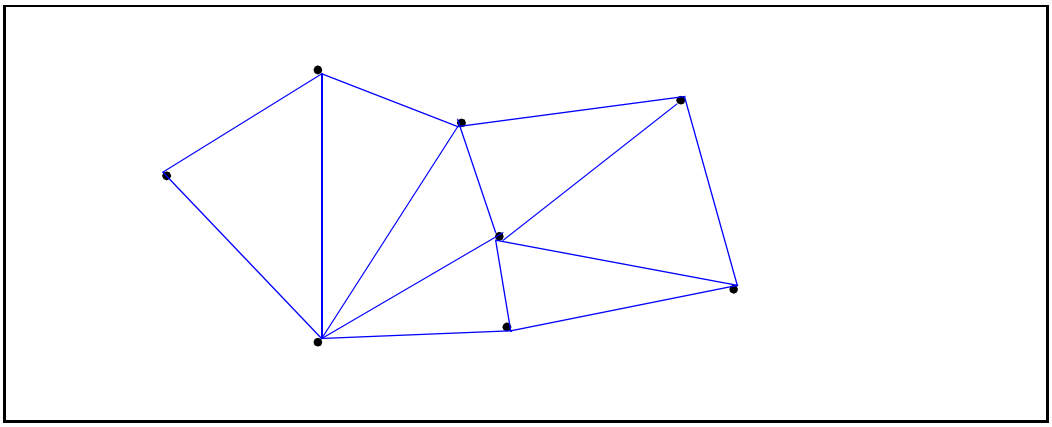


Abbildung 3.4: Delauny-Triangulation aus obigem Voronoi-Diagramm.

Es gibt verschiedene Verfahren aus dem Bereich “Computational Geometry”, um Voronoi-Diagramme für eine gegebene Menge von Ankerpunkten zu konstruieren. Man unterscheidet dabei nach Divide-and-Conquer-Algorithmus, Plane-Sweep-Konstruktion oder über den Umweg einer Delaunay-Triangulation. Für einen Überblick zu Voronoi-Diagrammen und zu ihrer Konstruktion siehe [Ottmann & Widmayer 96] oder den umfangreichen Survey von [Aurenhammer 91]. In [Schneider 97] wurde der Divide-and-Conquer-Algorithmus von [Shamos & Hoey 75] gewählt, der ein Voronoi-Diagramm der Dimension $d = 2$ mit k Ankerpunkten in $O(k \log k)$ Schritten berechnet. Höher-dimensionale Voronoi-Diagramme können z.B. erzeugt werden durch Transformation der entsprechenden höher-dimensionalen konvexen Hülle. Für einen Überblick über solche einbettenden Techniken siehe Aurenhammer ([Aurenhammer 91], Paragraph 1.3.4).

Voronoi-Regionen bilden keine (multivariaten) Histogramme im Sinne der Definition von Abschnitt 2.3.2. Dies liegt daran, daß die Voronoi-Regionen nicht paarweise disjunkt sind, da sie sich an den Rändern überdecken können. Trotzdem läßt sich der allgemeine Histogrammselektivitätsschätzer auf die Voronoi-Regionen zur Selektivitätsschätzung anwenden. Dazu werden alle Stichprobenelemente X_i entsprechend ihrer Zugehörigkeit zu den k Voronoi-Regionen C_j gewichtet und den Voronoi-Regionen, in denen sie liegen, entsprechend des Gewichtes zugerechnet. Liegt z.B. das Stichprobenelement X_i auf dem Rand der beiden Voronoi-Regionen C_j und C_l und in keiner anderen Region, so wird es jedem dieser beiden “Bins” C_j und C_l mit dem Wert 0,5 zugerechnet und allen anderen mit dem Wert 0. Im Unterschied zur Definition des allgemeinen Histogrammschätzers kann die Anzahl n_j der Elemente eines Bins daher auch rationale Werte annehmen. Will man dies vermeiden, so besteht eine Alternative darin, das Element zufällig nur genau einer der betreffenden Voronoi-Regionen zukommen zu lassen. Dieses Vorgehen ist in [Schneider 97] gewählt. Nun kann die Selektivität einer Anfrage mittels Gleichung (3.4) aus der Definition in Abschnitt 3.1 bestimmt werden. Diese Überlegungen seien unter Verwendung der ersten Variante in einer Definition zusammengefaßt:

Definition 3.7 (*Voronoi-Selektivitätsschätzer*):

Sei R eine Relation der Stelligkeit d, I eine Instanz von R und $X_1,...,X_n$ eine einfache Zufallsstichprobe der Größe n aus der Instanz. Desweiteren sei eine Anfrage $Q = Q(a,b)$ und k Ankerpunkte gegeben, die zu einem Voronoi-Diagramm mit k Voronoi-Regionen C_i, $i = 0...k\text{-}1$, führen. Die Stelligkeit $\lambda(X_j)$ eines Stichprobenelements X_j sei definiert als die Anzahl der C_i, die X_j enthalten:

$$\lambda(X_j) = \left|\{C_i : X_j \in C_i, i = 0 \ldots k-1\}\right|, j = 1...n \text{ mit } 0 < \lambda(X_j) \leq k. \tag{3.16}$$

Dann ist der *Voronoi-Selektivitätsschätzer* definiert als

$$\hat{\sigma}_V(a, b) = \frac{1}{n} \cdot \sum_{i=0}^{k-1} \frac{\tilde{n}_i}{Vol(C_i)} \cdot \psi_i(Q) \text{ wobei } \tilde{n}_i = \sum_{j=1}^{n} \frac{I_i(X_j)}{\lambda(X_j)} \tag{3.17}$$

und $Vol(C_i)$ und $\psi_i(Q)$ wie in Definition 3.2.

Bemerkung 3.5:

Es ist leicht zu sehen, daß $\sum_{i=0}^{k-1} \tilde{n}_i = \sum_{i=0}^{k-1} \sum_{j=1}^{n} \frac{I_i(X_j)}{\lambda(X_j)} = \sum_{j=1}^{n} \sum_{i=0}^{k-1} \frac{I_i(X_j)}{\lambda(X_j)} = \sum_{j=1}^{n} 1 = n.$

Im allgemeinen gilt sowohl $\tilde{n}_i \neq \tilde{n}_j$ für $i \neq j$ als auch $Vol(C_i) \neq Vol(C_j)$ für $i \neq j$, d.h. eine Vereinfachung in Analogie zum Equi-Width oder Equi-Depth-Histogrammschätzer ist nicht ohne weiteres gegeben (vgl. jedoch unten zur Wahl der Ankerpunkte).

Eine weiterhin offene Frage ist die Wahl der Ankerpunkte. Liegen die Ankerpunkte z.B. auf einem äquidistanten Gitter, so ergeben sich abgesehen vom Rand Voronoi-Regionen gleicher Größe und Gestalt, siehe Abbildung 3.5. Die so entstehenden Voronoi-Diagramme seien *Equi-Width-Voronoi-Diagramme* genannt. Eine andere einfache Methode zur Wahl der Ankerpunkte besteht darin, aus der Instanz *I* eine einfache Zufallsstichprobe $a_0,...,a_{k-1}$ der Größe *k* zu erzeugen und ihre Elemente als Ankerpunkte zu definieren. Eine weitere Möglichkeit besteht darin, die Ankerpunkte so zu wählen, daß sie einer vorgegebenen Verteilung folgen. Im folgenden wird eine Methode zur Bestimmung der Ankerpunkte (und Voronoi-Regionen) vorgeschlagen, wobei diese nach gewissen Kriterien optimiert werden.

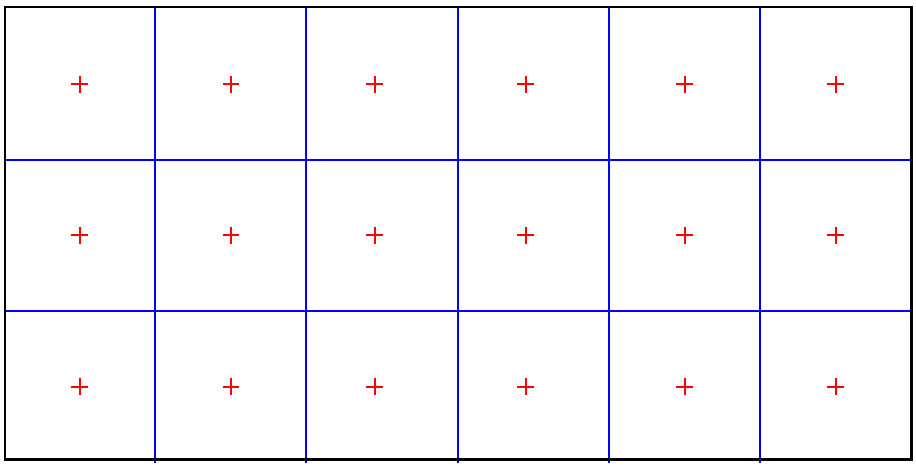


Abbildung 3.5: Equi-Width-Voronoi-Diagramm zu 18 äquidistant verteilten Ankerpunkten.

In [Schreiber 91] wurde ein inkrementelles Verfahren vorgeschlagen, welches eine Voronoi-Partitionierung nach vorgegebenen Kriterien zur Erzielung einer k-means ähnlichen Clusterung berechnet. K-means ist ein in der Statistik wohlbekannter Clusteralgorithmus mit einer vorgegebenen Kostenfunktion ([Hartigan 75]). Das Verfahren von [Schreiber 91] benutzt dazu die folgende zu minimierende gewichtete Kostenfunktion:

$$S = \sum_{j=1}^{k} s_j \text{ mit } s_j = \left(\sum_{i \in I_j} w_i \left\| X_i - a_j \right\|^2 \right) \Big/ \left(\sum_{i \in I_j} w_i \right), \tag{3.18}$$

wobei $I_j = \{ i = 1 \ldots n \mid X_i \in V(a_j) \}$, $0 < w_i \in \Re$ und $\| . \|$ die euklidische Norm.

Als notwendige Voraussetzung für die Minimierung muß für die partiellen Ableitungen $\frac{\partial s_j}{\partial a_j} = 0$ und $\frac{\partial^2 s_j}{\partial a_j} \geq 0$ gelten ([Schreiber 91]).

Im Unterschied zur k-means Clusterung beginnt der iterative Algorithmus von [Schreiber 91] mit einer einzigen Voronoi-Region V_1, die die gesamte Domäne umfaßt. Dieser Ansatz hat den Vorteil, daß die starke Abhängigkeit des k-means Verfahren von einer "guten" Anfangsclusterung entfällt. In jedem Iterationsschritt t wird anhand eines Kostenkriteriums bestimmt, welche der t Voronoi-Regionen "geteilt" werden soll. In [Schreiber 91] wird als Kostenkriterium die Fehlerfunktion (3.18) gewählt.

Bei einer "Teilung" werden die Stichproben der betrachteten Voronoi-Region $V_j = V_j(t)$ nach einem Teilungskriterium in zwei Cluster C_1 und C_2 unterteilt. Die Zentren dieser beiden Cluster werden jeweils als gewichtete Durchschnitte aller in dem jeweiligen Cluster enthaltenen Stichprobenwerte berechnet. Sie definieren gleichzeitig die beiden neuen Ankerpunkte $a_j(t+1)$ und $a_{t+1}(t+1)$, wobei der alte Ankerpunkt $a_j(t)$ der betrachteten Voronoi-Region $V_j(t)$ verschwindet:

$$a_j = \left(\sum_{i \in \{i | X_i \in C_1\}} w_i X_i \right) \Big/ \left(\sum_{i \in \{i | X_i \in C_1\}} w_i \right) \text{ und}$$

$$a_{t+1} = \left(\sum_{i \in \{i | X_i \in C_2\}} w_i X_i \right) \Big/ \left(\sum_{i \in \{i | X_i \in C_2\}} w_i \right).$$

Anhand der neuen Ankerpunkte a_1, ..., a_{t+1} lassen sich die neuen Voronoi-Regionen $V_1(t+1)$, ..., $V_{t+1}(t+1)$ bestimmen. Dabei ändern sich höchstens die der ursprünglichen Voronoi-Region $V_j(t)$ direkt benachbarten Voronoi-Regionen, d.h. höchstens die Voronoi-Regionen, die mit dem ursprünglichen $V_j(t)$ eine gemeinsame Kante besitzen. Die ursprüngliche Voronoi-Region $V_j(t)$ wird durch die beiden neu gebildeten Voronoi-Regionen $V_j(t+1)$ und $V_{t+1}(t+1)$ ersetzt. Das Updaten der Voronoi-Regionen erfordert also in jedem Iterationsschritt lediglich lokale Berechnungen.

[Schreiber 91] schlägt als Teilungskriterium vor, daß die Voronoi-Region $V_j(t)$ durch eine Hyperebene orthogonal zu derjenigen Achse p mit der größten projektiven Varianz in der betrachteten Voronoi-Region $V_j(t)$ geteilt wird, wobei die Hyperebene durch den ursprünglichen Ankerpunkt $a_j(t)$ geht:

$$p: \sum_{i \in I_j} w_i (X_{i,p} - a_{j,p}) = \max \left\{ \sum_{i \in I_j} w_i (X_{i,l} - a_{j,l}) \right\}$$

$$C_1 = \{X_i \,|\, i \in I_j \wedge X_{i,p} \leq a_{j,p}\} \text{ und } C_2 = \{X_i \,|\, i \in I_j \wedge X_{i,p} > a_{j,p}\}$$

Der Algorithmus wird solange fortgesetzt bis ein definiertes Abbruchkriterium erreicht ist. Sinnvolles Abbruchkriterium kann z.B. eine vorgegebene Anzahl *k* von Voronoi-Regionen oder die Unterschreitung eines vorgegebenen Fehlermaßes sein. Im letzteren Fall muß somit im Gegensatz zum k-means Verfahren die Anzahl der Cluster nicht von vorneherein bekannt sein. Das so entstehende Voronoi-Diagramm kann als Grundlage für die Voronoi-Selektivitätsschätzung wie in Definition 3.7 verwendet werden.

Vorteil des Verfahrens zur Selektivitätsschätzung besteht darin, daß die Restriktion einer rechteckigen Partitionierung aufgehoben wird. Des weiteren kann eine weitere Kostenfunktion integriert werden bei der Bestimmung, welche Voronoi-Region geteilt werden soll. Hierzu gehört z.B. die Forderung, daß die Voronoi-Regionen eine nahezu gleiche Anzahl von Elementen besitzen sollen. Nachteil des Verfahrens ist der höhere Rechenaufwand bei der Selektivitätsschätzung zur Bestimmung der Schnittmenge von Anfragebereich und Voronoi-Region. Zudem ist nicht klar, inwieweit die Verteilungen innerhalb der jeweiligen Voronoi-Regionen zu einer guten Selektivitätsschätzung beitragen.

Das beschriebene Verfahren zur Selektivitätsschätzung mittels Voronoi-Diagrammen, die durch den Algorithmus von [Schreiber 91] bestimmt werden, wurde in [Schneider 97] ausführlich untersucht und um die Einbeziehung von Nachbarschaftseigenschaften erweitert. [Schneider 97] vergleicht das Verfahren im bivariaten Fall mit der direkten Selektivitätsschätzung aus der Stichprobe und erzielt damit bessere Ergebnisse bei gleichverteilten, korrelierten und realen Testdaten.

Im folgenden ist der Algorithmus zur Voronoi-Selektivitätsschätzung beschrieben:

Algorithmus 3.3 (*Voronoi-Selektivitätsschätzer*):

A. Bilden der Voronoi-Klassen:

Gegeben: Dimension *d*, eine Relation *R* mit Instanz *I* und Stichprobe $X_1,...,X_n$ der Größe *n*.

1. Erzeuge Voronoi-Diagramm z.B. mittels dem adaptiven Voronoi-basiertem *k*-means-artigem Clusterverfahren von [Schreiber 91]. Die *k* Voronoi-Regionen partitionieren den Datenraum in Voronoi-Klassen C_i, $i = 0...k\text{-}1$, *k* vorgegeben.

2. a) Für alle Stichprobenelemente X_j berechne $\lambda(X_j)$ mittels (3.16).

   b) Berechne nun $\tilde{n}_j$ aus (3.17).

2. Für jede Voronoi-Klasse C_i, $i = 0...k\text{-}1$, speichere das zugehörige $\tilde{n}_j$.

B. Bearbeiten der Anfrage:

Gegeben: wie A. mit entsprechendem Histogramm $(C_j, \tilde{n}_j)_{j=1...k}$ sowie eine Bereichsanfrage $Q(a,b)$.

1. Die Selektivitätsschätzung ergibt sich nun durch Anwendung des Voronoi-Selektivitätsschätzers gemäß Gleichung (3.17) definiert.

3.3 Selektivitätsschätzung mittels Average Shifted Histogrammen

Die Nachteile des Histogrammschätzers zur Dichteschätzung wurden bereits in Kapitel 2.3.3 aufgezeigt und treffen gleichermaßen bei der Selektivitätsschätzung zu. Es handelt sich im Wesentlichen um die analytische Unstetigkeit der Schätzfunktion, die starke Abhängigkeit des Schätzers von der Wahl des Anfangspunktes und die schlechte Konvergenzrate. Zur Vermeidung der Schwächen wurden in Kapitel 2.3.3 verschiedene Erweiterung des Histogrammschätzers vorgestellt. Diese lassen sich auch zur Selektivitätsschätzung anwenden. Dabei beschränkt sich diese Arbeit auf die Verwendung des Average Shifted Histogrammschätzers zur Selektivitätsschätzung.

Zur Herleitung einer Definition des Average Shifted Histogrammselektivitätsschätzers genügt es die entsprechenden Gleichungen - (2.38) bzw. (2.39) - in Gleichung (2.8) einzusetzen.

$$\hat{\sigma}_{ASH}(a,b) = \int_a^b \hat{f}_{ASH}(t)dt = \int_a^b \frac{1}{m}\sum_{j=1}^{m} \hat{f}_j(x)dt = \frac{1}{m}\sum_{j=1}^{m}\int_a^b \hat{f}_j(x)dx\,.$$

Dies erlaubt die folgende Definition:

Definition 3.8 (*univariater Average Shifted Histogramm Selektivitätsschätzer*):

Sei R eine Relation der Stelligkeit $d = 1$, I eine Instanz von R und $X_1,...,X_n$ eine einfache Zufallsstichprobe der Größe n aus der Instanz sowie $\hat{f}_1,...,\hat{f}_m$ eine Menge von m Equi-Width-Histogrammschätzern mit gleicher Binweite h aber unterschiedlichen Startpunkten $x_0 = 0, h/m, ..., (m-1)h/m$. Desweiteren sei eine Anfrage $Q = Q(a,b)$ gegeben. Dann ist der *univariate Average Shifted Histogramm Selektivitätsschätzer* definiert als:

$$\hat{\sigma}_{ASH}(a,b) = \frac{1}{m}\sum_{j=1}^{m}\int_a^b \hat{f}_j(x)dx \tag{3.19}$$

Im multivariaten Fall ergibt sich ebenfalls durch Einsetzen:

$$\begin{aligned}\hat{\sigma}_{ASH}(a,b) &= \int_a^b \frac{1}{m_1 \ldots m_d} \sum_{j_1=1}^{m_1} \cdots \sum_{j_d=1}^{m_d} \hat{f}_{j_1 \ldots j_d}(x)dt \\ &= \frac{1}{m_1 \ldots m_d} \sum_{j_1=1}^{m_1} \cdots \sum_{j_d=1}^{m_d} \int_a^b \hat{f}_{j_1 \ldots j_d}(x)dx \end{aligned} \tag{3.20}$$

Man beachte den erhöhten Rechenaufwand mit steigender Größe der Dimension. Bereits im bivariaten Fall müssen $m_1 \cdot m_2$ Histogramme berechnet werden. Dies ist der Hauptnachteil, der den ASH-Selektivitätsschätzer für praktische multivariate Anwendungen ungeeignet macht.

Der ASH-Selektivitätsschätzer wurde in Kapitel 5 auf sämtliche Testdatensätze angewendet. Die Ergebnisse sind dort dokumentiert.

3.4 Selektivitätsschätzung mittels Kernfunktionen

Kernschätzer besitzen gegenüber Histogrammschätzern und Average Shifted Histogrammschätzern u.a. den Vorteil, daß sie eine glatte Schätzung der Dichtefunktion ohne Benutzung apriori berechneter Histogramme liefern. Sie werden daher in dieser Arbeit zur Kernselektivitätsschätzung angewendet.

3.4.1 Kernselektivitätsschätzer

In diesem Kapitel wird der Kernschätzer zur Schätzung der Selektivität einer Bereichsanfrage erweitert. Dazu genügt es, Gleichung (2.54) mit der Definition des Kerndichteschätzers in Gleichung (2.5) zur Selektivitätsschätzung einzusetzen:

$$\hat{\sigma}_K(a,b) = \int_{[a,b]} \hat{f}_K(x)dx = \int_{[a,b]} \frac{1}{n} \cdot \sum_{i=1}^{n} K_H(x - X_i)dx = \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{[a,b]} K_H(x - X_i)dx .$$

Definition 3.9 (*Kernselektivitätsschätzer*):

Sei R eine Relation, I eine Instanz von R und $X_1,...,X_n$ eine einfache Zufallsstichprobe der Größe n aus der Instanz. Desweiteren sei eine Anfrage $Q = Q(a,b)$ gegeben. Dann ist der *Kernselektivitätsschätzer* definiert als:

$$\hat{\sigma}_K(a,b) = \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{[a,b]} K_H(x - X_i)dx \text{ mit } K_H(x) = |H|^{-1} K(H^{-1}x) \qquad (3.21)$$

und der Bandbreitenmatrix H wie in Definition 2.19 gegeben.

Bemerkung 3.6:

- Im univariaten Fall ergibt sich für Gleichung (3.21) mit Gleichung (2.55):

$$\hat{\sigma}_K(a,b) = \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{a}^{b} \frac{1}{h} \cdot K\left(\frac{x - X_i}{h}\right)dx \qquad (3.22)$$

  Ist h sogar unabhängig von x und X_i, d.h. h = const., so läßt sich der Faktor $1/h$ ganz bis vor die Summe herausziehen.

- Ist die Bandbreiten-Matrix eine Diagonalmatrix $H = \text{diag}(h_1^2,...h_d^2)$ mit $h_j \in \Re^{>0}$

gegeben, so schreibt sich Gleichung (3.21) mit Gleichung (2.56):

$$\hat{\sigma}_K(a,b) = \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{[a,b]} \frac{1}{\prod_{j=1}^{d} h_j} \cdot K\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_d} \right) dx \tag{3.23}$$

Sind in der obigen Diagonalmatrix alle h_j gleich einem konstantem h, so vereinfacht sich Gleichung (3.23) weiter zu:

$$\hat{\sigma}_K(a,b) = \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{[a,b]} \frac{1}{h^d} K\left(\frac{1}{h^d}(x - X_i) \right) dx \tag{3.24}$$

- Im bivariaten Fall ergibt sich für Gleichung (3.21) mit $H = diag(h_1^2, h_2^2)$:

$$\hat{\sigma}_K(a,b) = \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{a_2}^{b_2} \int_{a_1}^{b_1} \frac{1}{h_1 h_2} \cdot K\left(\frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2} \right) dx_1 dx_2 \tag{3.25}$$

Im folgenden wird aufgrund praktischer Überlegungen davon ausgegangen, daß es sich bei der Bandbreiten-Matrix um eine Diagonal-Matrix $H = \text{diag}(h_1^2, ..., h_d^2)$ handelt.

Sei im weiteren zunächst der univariate Fall mit $d = 1$ und Bandbreite h behandelt. Bereits in Kapitel 2.3.4 wurden die Vorzüge der Epanechnikow-Kernfunktion sowohl bzgl. mathematischer Eigenschaften als auch wegen der Einfachheit der Berechnung mit Computeralgorithmen herausgestellt. Im folgenden wird daher als Kernfunktion der Epanechnikow-Kern verwendet. Da der Epanechnikow-Kern die Eigenschaften einer Dichtefunktion besitzt, ist es zunächst sinnvoll als Stammfunktion eine passende Funktion mit den Eigenschaften einer Verteilungsfunktion zu wählen. Dazu eignet sich die folgende Stammfunktion F_{Epa}, vgl. auch Abbildung 3.6:

$$F_{Epa}(t) = \frac{1}{2} + \begin{cases} \frac{1}{4} t(3 - t^2) & \text{falls } |t| \le 1 \\ \frac{1}{2} \text{sgn}(t) & \text{falls } |t| > 1 \end{cases} \tag{3.26}$$

Bei der Berechnung des bestimmten Integrales $\int_a^b K(x)dx = F(b) - F(a)$ kann der konstante Summand 1/2 entfallen.

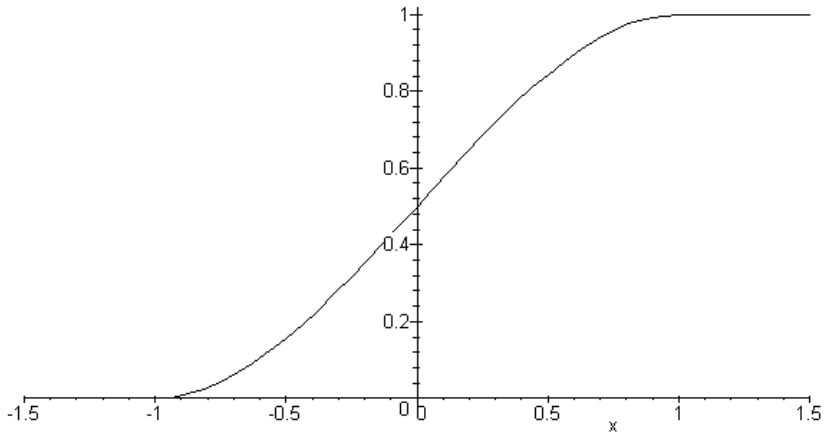


Abbildung 3.6: Stammfunktion des Epanechnikow-Kerns

Bei Anwendung von Gleichung (3.22) werden Kernfunktionen um die Stichprobenelemente mit der Bandbreite h gebildet. Durch Substitution von $t = (x - X_i) / h$ werden diese auf den normierten Kern mit Support $[-1,1]$ abgebildet (vgl. Abbildung 3.7).

$$\hat{\sigma}_K(a, b) = \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{(a - X_i)/h}^{(b - X_i)/h} K(t)dt \tag{3.27}$$

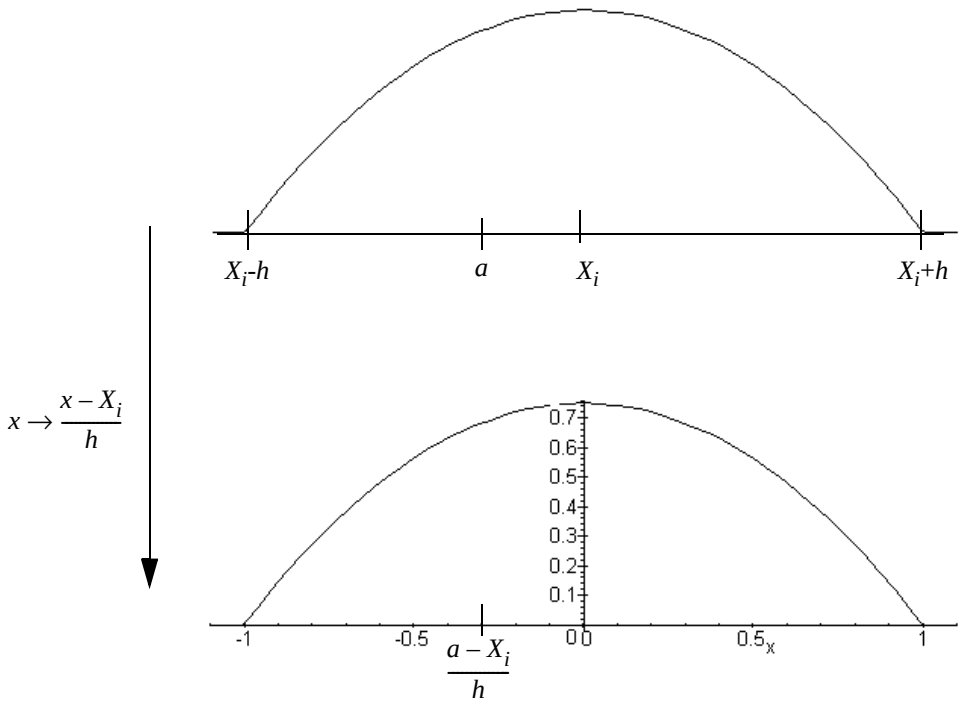


Abbildung 3.7: Transformation der Kernfunktion auf den Support $[-1,1]$.

Abbildung 3.8 veranschaulicht weiterhin, daß das Integral 0 ergibt, sobald der Support der Kernfunktion über einem Stichprobenelement ganz außerhalb des Intervalls $[a - h,b + h]$ liegt, und 1 ist, sobald der Support ganz innerhalb des Intervalls $[a + h,b - h]$ liegt.

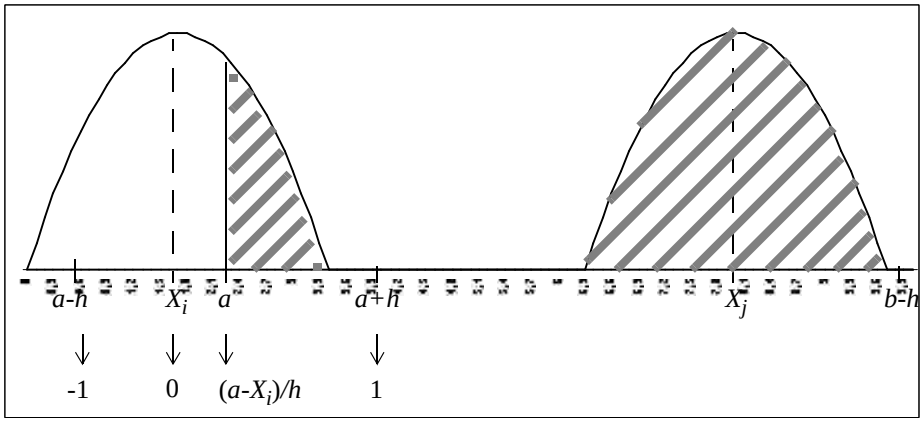


Abbildung 3.8: Integration der Kernfunktion zur Selektivitätsschätzung bei gegebener Anfrage $Q(a,b)$

Für den Epanechnikow-Kern gilt die folgende Fallunterscheidung:

$$\int_{(a-X_i)/h}^{(b-X_i)/h} K(t)dt = \begin{cases} 0 \text{ , falls } b + h \le X_i \vee a - h \ge X_i \\ 1 \text{ , falls } b - h \ge X_i \ge a + h \end{cases} \tag{3.28}$$

Lediglich für diejenigen Stichprobenwerte, bei denen der Support der zugehörigen Kernfunktion eine echte nicht-leere Schnittmenge mit dem Anfragebereich bildet, muß das Integral explizit ausgewertet werden (vgl. Abbildung 3.8). Dies trifft i.a. nur auf einen sehr kleinen Teil der Stichprobenwerte zu. Der Einfluß der Lage eines Stichprobenelementes auf die Kernselektivitätsschätzung wird in Abbildung 3.9 anschaulich gemacht unter Verwendung des Epanechnikow-Kerns. Insbesondere bedeutet dies einen großen Vorteil für Berechnungen mit Computern, da nur wenig berechnet werden muß. Für den Algorithmus bedeutet dies, daß Stichprobenelemente außerhalb des Intervalls $[a - h,b + h]$ nicht berücksichtigt werden und Stichprobenelemente innerhalb des Intervalls $[a + h,b - h]$ lediglich gezählt werden. Für die Stichprobenelemente im Intervall $[a - h,a + h]$ und $[b - h,b + h]$ muß

$$F_{Epa}\left(\frac{b - X_i}{h}\right) - F_{Epa}\left(\frac{a - X_i}{h}\right) \tag{3.29}$$

explizit mit Gleichung (3.26) ausgewertet werden.

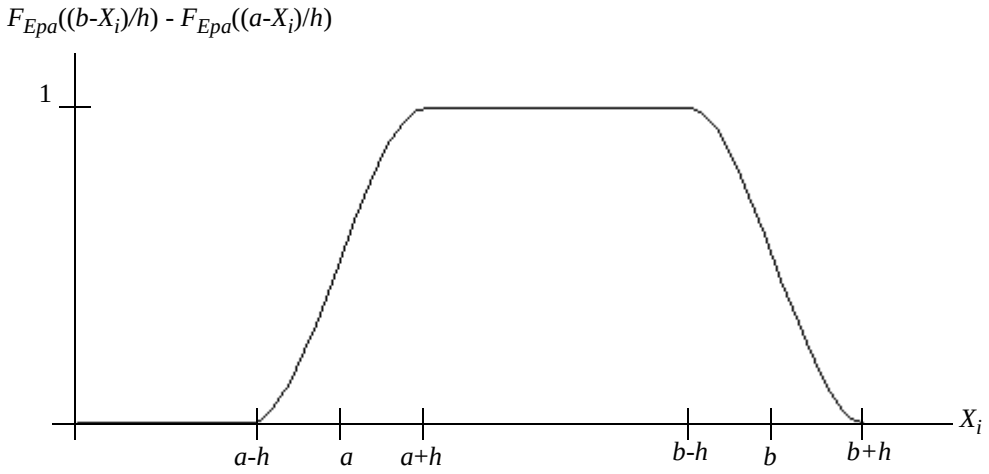


Abbildung 3.9: Einfluß eines möglichen Stichprobenelementes X_i auf die Selektivitätsschätzung bei gegebener Anfrage $Q(a,b)$.

Dies wird in folgendem Algorithmus zusammengefaßt:

Algorithmus 3.4 (*univariate Kernselektivitätsschätzung*):

Gegeben: Dimension d=1, eine Relation R mit Instanz I und Stichprobe $X_1,...,X_n$ der Größe n, Bereichsanfrage $Q(a,b)$, Epanechnikow-Kernfunktion K mit Bandbreite h.

1. Setze S gleich der Anzahl der Stichprobenelemente X_i im Intervall $[a + h,b – h]$.
2. Berechne für alle Stichprobenelemente X_i aus $[a – h,a + h]$ und aus $[b – h,b + h]$ das Integral aus Gleichung (3.29) und addiere jeweils den Wert zu S hinzu.
3. Die geschätzte Selektivität der Anfrage $Q(a,b)$ ergibt sich nun als $\hat{\sigma}(Q) = S/n$.

Man beachte, daß hier im Gegensatz zu den Histogramm-Selektivitätsschätzern aus Kapitel 3.2 keine Informationen apriori berechnet werden müssen.

Die oben getroffene Fallunterscheidung läßt sich auf den multivariaten Fall verallgemeinern. Der Einfachheit halber sei hier nur der bivariate Fall mit (Epanechnikow-)Produktkern und Diagonalbandbreitenmatrix $H = diag(h_1,h_2)$ beschrieben.

Da mit dem Satz von Fubini gilt:

$$F_{Epa}^{P}(t) = \int\int K(t_1)K(t_2)dt_1 d(t_2) = \int\int K(t_1)dt_1 K(t_2)dt_2 \quad , \qquad (3.30)$$
$$= \int F_{Epa}(t_1)K(t_2)dt_2 = F_{Epa}(t_1)F_{Epa}(t_2)$$

ergibt sich als eine Stammfunktion des bivariaten Epanechnikow-Produktkerns (2.50), vgl. Abbildung 3.10:

$$F_{Epa}^{P}(t) = \frac{1}{4} + \begin{cases} (3t_1 - {t_1}^3)(3t_2 - {t_2}^3)/16 & , \text{falls } |t_1| \le 1 \wedge |t_2| \le 1 \\ 3/4 & , \text{falls } t_1 > 1 \wedge t_2 > 1 \\ -1/4 & , \text{falls } t_1 < -1 \vee t_2 < -1 \\ (3t_1 - {t_1}^3)/8 & , \text{falls } |t_1| \le 1 \wedge t_2 > 1 \\ (3t_2 - {t_2}^3)/8 & , \text{falls } |t_2| \le 1 \wedge t_1 > 1 \end{cases} \qquad (3.31)$$

Auch im bivariaten Fall hat die Stammfunktion des Epanechnikow-Kerns wieder die Eigenschaften einer Verteilungsfunktion.

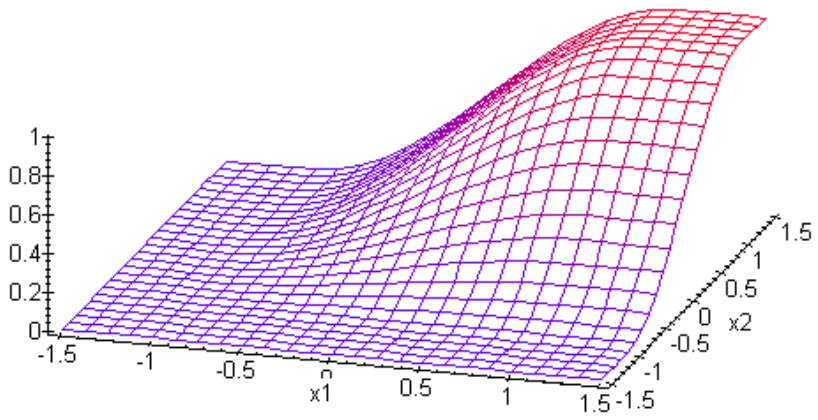


Abbildung 3.10: Graph der Stammfunktion des bivariaten Epanechnikow-Kerns

Beim Einsetzen der Kernfunktion in den Selektivitätsschätzer ergeben sich Terme der Form

$$\int_{a_2}^{b_2}\int_{a_1}^{b_1} K\left(\frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2}\right) dx_1 dx_2$$

Dabei werden die Kerne um X_i betrachtet. Durch Substitution von $t = (t_1, t_2)^t = \left(\frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2}\right)^t$ ergeben sich Terme mit Kernen um den Nullpunkt:

$$\int_{a_2}^{b_2}\int_{a_1}^{b_1} K\left(\frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2}\right) dx_1 dx_2 = \int_{\frac{a_2 - X_{i2}}{h_2}}^{\frac{b_2 - X_{i2}}{h_2}} \int_{\frac{a_1 - X_{i1}}{h_1}}^{\frac{b_1 - X_{i1}}{h_1}} h_1 h_2 K(t) dt_1 dt_2 . \qquad (3.32)$$

Damit ergibt sich für den bivariaten Selektivitätsschätzer mit Epanechnikow-Kern:

$$\hat{\sigma}_K(a, b) = \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{\frac{a_2 - X_{i2}}{h_2}}^{\frac{b_2 - X_{i2}}{h_2}} \int_{\frac{a_1 - X_{i1}}{h_1}}^{\frac{b_1 - X_{i1}}{h_1}} K(t) dt_1 dt_2 . \qquad (3.33)$$

Auch im multivariaten Fall ist das Integral der Epanechnikow-Kernfunktion über einem Stichprobenelement X_i gleich 1, wenn X_i innerhalb des d-dim. Intervalls $[a + h, b - h]$ liegt, und gleich 0, wenn es außerhalb des Intervalls $[a - h, b + h]$ liegt. Zur einfacheren Schreibweise seien die folgenden Bezeichnungen im zweidimensionalen Fall eingeführt:

$$\begin{aligned}
I(a, b) &= [a + h, b - h] \qquad , \qquad (3.34)\\
A(a, b) &= \overline{[a - h, b + h]}\\
R_1(a, b) &= [(a_1 - h_1, a_2 + h_2)^t, (a_1 + h_1, b_2 - h_2)^t]\\
&\quad \cup [(b_1 - h_1, a_2 + h_2)^t, (b_1 + h_1, b_2 - h_2)^t]\\
R_2(a, b) &= [(a_1 + h_1, a_2 - h_2)^t, (b_1 - h_1, a_2 + h_2)^t]\\
&\quad \cup [(a_1 + h_1, b_2 - h_2)^t, (b_1 - h_1, b_2 + h_2)^t]\\
E(a, b) &= [(a_1 - h_1, a_2 - h_2)^t, (a_1 + h_1, a_2 + h_2)^t]\\
&\quad \cup [(a_1 - h_1, b_2 - h_2)^t, (a_1 + h_1, b_2 + h_2)^t]\\
&\quad \cup [(b_1 - h_1, a_2 - h_2)^t, (b_1 + h_1, a_2 + h_2)^t]\\
&\quad \cup [(b_1 - h_1, b_2 - h_2)^t, (b_1 + h_1, b_2 + h_2)^t]
\end{aligned}$$

vgl. hierzu Abbildung 3.11

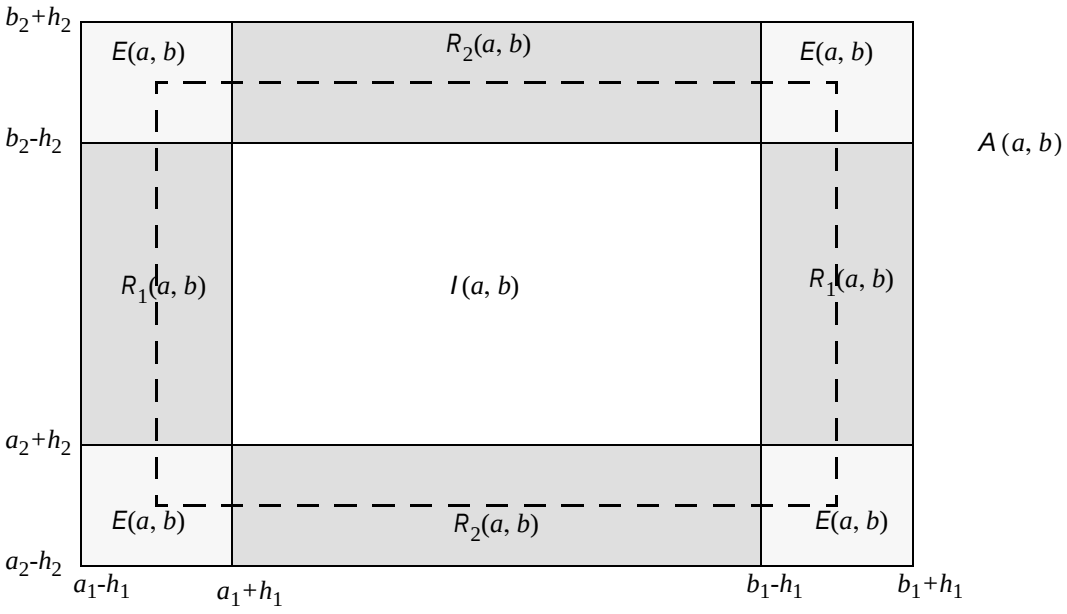


Abbildung 3.11: Zweidimensionaler Anfragebereich $[a,b]$ (Fläche innerhalb gestrichelter Linie) mit Bandbreite $H = diag(h_1^2, h_2^2)$.

Dies führt im bivariaten Fall zu folgenden Fallunterscheidungen:

$$\int_{\frac{a_2 - X_{i2}}{h_2}}^{\frac{b_2 - X_{i2}}{h_2}} \int_{\frac{a_1 - X_{i1}}{h_1}}^{\frac{b_1 - X_{i1}}{h_1}} K(t)dt_1 dt_2 = \begin{cases} 1 & , X_i \in I(a, b) \\ 0 & , X_i \in A(a, b) \\ F_{Epa}(t_1)\Big|_{(a_1 - X_{i1})/h_1}^{(b_1 - X_{i1})/h_1} & , X_i \in R_1(a, b) \\ F_{Epa}(t_2)\Big|_{(a_2 - X_{i2})/h_2}^{(b_2 - X_{i2})/h_2} & , X_i \in R_2(a, b) \end{cases} \tag{3.35}$$

Für $X_i \in E(a, b)$ muß das komplette Integral explizit berechnet werden (mittels Zeile 1 von Gleichung (3.31)). Die Konsequenzen für die praktische Berechnung sind wie im univariaten Fall gelagert. Insbesondere muß wieder nur ein geringer Anteil von Integralen explizit berechnet werden, die Integrale bzgl. der X_i, die in $I(a,b)$ liegen, werden lediglich gezählt, die Integrale bzgl. der X_i, die in $A(a,b)$ liegen, werden nicht berücksichtigt. Der Einfluß eines möglichen Stichprobenelementes X_i auf die Selektivitätsschätzung einer gegebenen Anfrage beim bivariaten Epanechnikow-Kern ist in Abbildung 3.12 dargestellt.

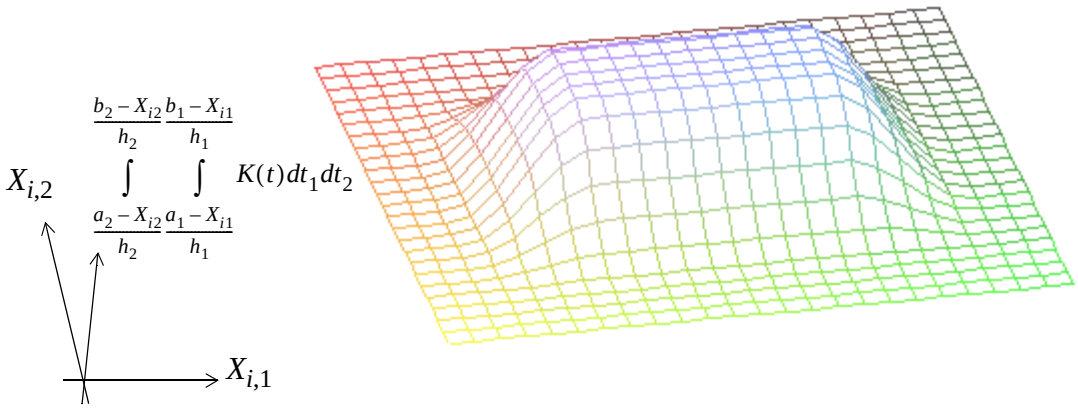


Abbildung 3.12: Einfluß eines möglichen Stichprobenelementes $X_i = (X_{i1}, X_{i2})^t$ auf die Kernselektivitätsschätzung im bivariaten Fall bei gegebener Anfrage $Q(a,b)$.

Die Erweiterung auf höherdimensionale Selektivitätsschätzung ist ohne weiteres möglich, allerdings sind deutlich mehr Fallunterscheidungen zu treffen.

Für den Algorithmus ergibt sich im bivariaten Fall:

Algorithmus 3.5 (*bivariate Kernselektivitätsschätzung*):

Gegeben: Dimension d=2, eine Relation R mit Instanz I und Stichprobe $X_1,...,X_n$ der Größe n, Bereichsanfrage $Q(a,b)$, Epanechnikow-Produktkern K mit Bandbreite $H = diag(h_1^2, h_2^2)$.

1. Setze S gleich der Anzahl der Stichprobenelemente X_i aus $I(a,b)$.
2. Berechne für alle Stichprobenelemente X_i aus $R(a,b)$ die zugehörigen Integrale gemäß Gleichung (3.35) und (3.26), sowie für X_i aus $E(a,b)$ das komplette Integral mittels Gleichung (3.31) und addiere die Werte zu S hinzu.
3. Die geschätzte Selektivität der Anfrage $Q(a,b)$ ergibt sich nun als $\hat{\sigma}(Q) = S/n$.

Kernfunktionen zur Selektivitätsschätzung wurden in dieser Arbeit ausführlich anhand univariater und bivariater, künstlicher und realer Datensätze getestet, und die Ergebnisse sind in Kapitel 5 repräsentiert.

3.4.2 Kernselektivitätsschätzer mit Randbehandlung

In Kapitel 2.3.5 wurde das Phänomen der Randprobleme aufgegriffen und ausführlich bei Kerndichteschätzern diskutiert. Experimente zur Selektivitätsschätzung mit verschiedenen Datenmengen zeigen, daß der Fehler bei der Kernselektivitätsschätzung ebenfalls am Rand stark

ansteigen kann. Abbildung 3.13 zeigt als Beispiel diesen Effekt bei univariaten gleichverteilten Testdaten (s. Kap. 5). Auf der *x*-Achse sind die Mittelpunkte der Anfrageintervalle und auf der *y*-Achse die Differenz von wahrer und geschätzter Selektivität als Fehler aufgetragen. Die Größe der 1.000 verwendeten Anfrageintervalle betrug 1% der Domäne, die Daten bestanden aus 100.000 gleichverteilten Testdaten, vgl. Kapitel 5.3.1. Die Bandbreite betrug 112.000. Die Fehlerkurve zeigt deutlich, daß der Fehler im mittleren Bereich $[l+h+(b-a)/2,\ r-h-(b-a)/2]$ = [117.243, 931.332] zwischen -131 und 157 liegt, während er im Randbereich auf bis zu 597 ansteigt. Unter Berücksichtigung des Vorzeichens wird ersichtlich, daß der Kernschätzer am Rand die Selektivität deutlich unterschätzt. Dies war zu erwarten aufgrund der fehlenden Masse am Rand.

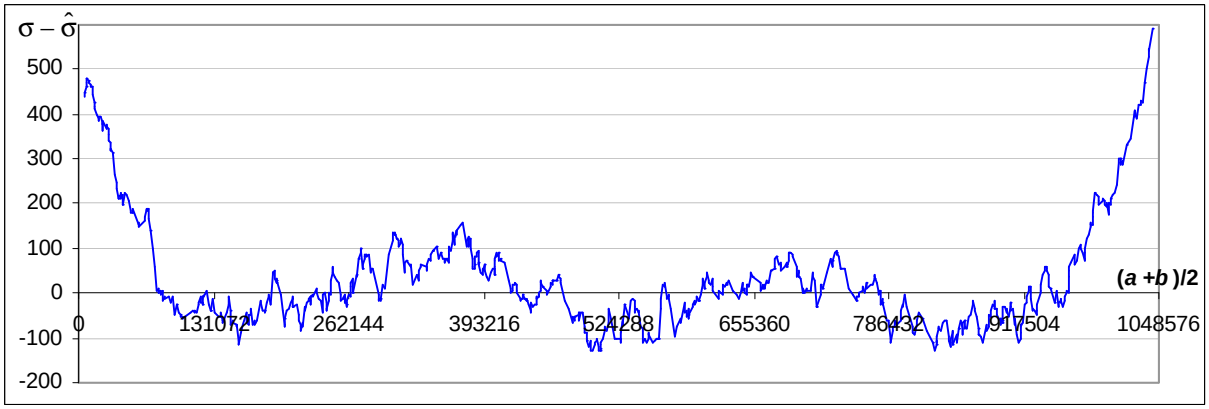


Abbildung 3.13: Randfehler bei Kernselektivitätsschätzung am Beispiel gleichverteilter Testdaten.

Die in Kapitel 2.3.5 diskutierten Ergebnisse zur Randbehandlung bei Kerndichteschätzern werden in diesem Abschnitt zur Selektivitätsschätzung angewendet. Dazu müssen die Randkerndichteschätzer in Analogie zu Kapitel 3.4.1 integriert werden. Dabei wird auf die Fallunterscheidung bei der Definition des Randkerns in Gleichung (2.81) Bezug genommen. Sei zunächst der univariate Fall betrachtet und eine Anfrage $Q(a,b)$ auf der Domäne $[l, r]$ gegeben ($l \le a < b \le r$), die Bandbreite h sei beliebig aber fest mit $h < (r-l)/2$. Dann sind folgende Fälle zu unterscheiden:

(i) $l + h \le a < b \le r - h$

(ii) $a < l + h \le b \le r - h$

(iii) $a < b < l + h$

(iv) $l + h \le a \le r - h < b$

(v) $r - h < a < b$

(vi) $a < l + h < r - h < b$

Die Fälle (iii) und (v) werden eher seltener auftreten, da nicht zu erwarten ist, daß die Anfragebreite kleiner ist als die Bandbreite *h*. Das gleiche gilt für Fall (vi), da i.a. sowohl die Anfragebreite als auch die Bandbreite sehr viel kleiner sind als die Domänenbreite. Im Fall (i) ist keine Randbehandlung erforderlich und die Selektivität wird wie in Gleichung (3.22) berechnet. Für die Fälle (ii) bis (vi) ist das Integral wie folgt aufzusplitten, wobei $q^l = (x-l)/h$ und $q^r = (r-x)/h$:

$$\text{(ii)}\quad \hat{\sigma}(a,b) = \frac{1}{nh}\cdot\sum_{i=1}^{n}\left(\int_a^{l+h} K^l\left(\frac{x-X_i}{h}, q^l\right)dx + \int_{l+h}^{b} K\left(\frac{x-X_i}{h}\right)dx\right)$$

$$\text{(iii)}\quad \hat{\sigma}(a,b) = \frac{1}{nh}\cdot\sum_{i=1}^{n}\int_a^b K^l\left(\frac{x-X_i}{h}, q^l\right)dx$$

$$\text{(iv)}\quad \hat{\sigma}(a,b) = \frac{1}{nh}\cdot\sum_{i=1}^{n}\left(\int_a^{r-h} K\left(\frac{x-X_i}{h}\right)dx + \int_{r-h}^{b} K^r\left(\frac{x-X_i}{h}, q^r\right)dx\right)$$

$$\text{(v)}\quad \hat{\sigma}(a,b) = \frac{1}{nh}\cdot\sum_{i=1}^{n}\int_a^b K^r\left(\frac{x-X_i}{h}, q^r\right)dx$$

$$\text{(vi)}\quad \hat{\sigma}(a,b) = \frac{1}{nh}\cdot\sum_{i=1}^{n}\left(\int_a^{l+h} K^l\left(\frac{x-X_i}{h}, q^l\right)dx + \int_{l-h}^{r-h} K\left(\frac{x-X_i}{h}\right)dx + \int_{r-h}^{b} K^r\left(\frac{x-X_i}{h}, q^r\right)dx\right)$$

Im folgenden werden die Renormalisierung, die Spiegelung und der Randkern von [Dong & Simonoff 94] aus Kapitel 2.3.5 auf die Selektivitätsschätzung mit Kernfunktionen angewendet.

Renormalisierung

Seien die Voraussetzungen wie in Kapitel 2.3.5 zur Renormalisierung gegeben. Sei weiterhin eine Anfrage $Q(a,b)$ gegeben. Dann ist der *renormalisierte Randkernselektivitätsschätzer* am Rand gegeben durch

$$\hat{\sigma}_R(a,b) = \int_a^b \hat{f}_R(x)dx = \int_a^b \frac{\hat{f}_K(x)}{a_0(q)}dx = \frac{1}{n}\sum_{i=1}^{n}\int_a^b \frac{K_h(x-X_i)}{a_0(q)}dx \qquad (3.36)$$

mit $a_0(q) = a_0^l(q) = \int_{-1}^{q} K(t)dt$, $x = l + qh$, $q \geq 0$ am linken Rand l und

mit $a_0(q) = a_0^r(q) = \int_{-q}^{1} K(t)dt$, $x = r - qh$, $q \geq 0$ am rechten Rand r.

Man beachte auch hier, daß $a_0(q) = 1$ für $q \geq 1$.

Spiegelung

Seien die Voraussetzungen wie in Kapitel 2.3.5 für die Spiegelung gegeben. Außerdem sei eine Anfrage $Q(a,b)$ gegeben. Der gespiegelte Randkerndichteschätzer $\hat{f}_S(x)$ bei gegebenen Rändern l und r ist nun definiert als

$$\hat{f}_S(x) = \frac{1}{n}\sum_{i=1}^{n}(K_h(x - X_i) + K_h(x + X_i - 2l) + K_h(x + X_i - 2r)) \tag{3.37}$$

Integration von a bis b bei gegebener Bereichsanfrage $Q(a,b)$ ergibt den *gespiegelten Selektivitätsrandkernschätzer*

$$\hat{\sigma}_S(a,b) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_a^b K_h(x - X_i)dx + \int_a^b K_h(x + X_i - 2l)dx + \int_a^b K_h(x + X_i - 2r)dx\right) \tag{3.38}$$

Substitution mit $t = (x - X_i)/h$ bzw. $t = (x + X_i - 2l)/h$ bzw. $t = (x + X_i - 2r)/h$ führt zu:

$$\hat{\sigma}_S(a,b) = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\frac{a-X_i}{h}}^{\frac{b-X_i}{h}} K(t)dt + \int_{\frac{a+X_i-2l}{h}}^{\frac{b+X_i-2l}{h}} K(t)dt + \int_{\frac{a+X_i-2r}{h}}^{\frac{b+X_i-2r}{h}} K(t)dt\right) \tag{3.39}$$

Man beachte, daß die zusätzlichen Stichprobenelemente für die Selektivitätsschätzung nur interessant sind, falls für eine gegebene Anfrage $Q(a,b)$ relevante Stichprobenelemente in den Randbereich fallen, d.h. im univariaten Fall für alle Anfragen mit $a < l + 2h$ oder $b > r - 2h$.

Algorithmus 3.6 (*univariate Kernselektivitätsschätzung mit Spiegelung*):

Gegeben: Dimension d=1, eine Relation R mit Instanz I und Stichprobe $X_1,...,X_n$ der Größe n, Bereichsanfrage $Q(a,b)$ auf dem Intervall $[l,r]$, Epanechnikow-Kernfunktion K mit Bandbreite h.

1. Setze S gleich der Anzahl der Stichprobenelemente X_i im Intervall $[a + h, b - h]$.

2. Berechne für alle Stichprobenelemente X_i aus $[a-h,a+h]$ und aus $[b-h,b+h]$ das Integral aus Gleichung (3.29) und addiere den Wert zu S hinzu.

3. Falls $a < l + 2h$ oder $b > r - 2h$: berechne ebenso für alle Punkte $2l - X_i$ und $2r + X_i$ mit linkem Rand l und rechtem Rand r, die noch in den Intervallen $[a-h,a+h]$ oder $[b-h,b+h]$ liegen, das Integral aus Gleichung (3.29) und addiere den Wert zu S hinzu.

3. Die geschätzte Selektivität der Anfrage $Q(a,b)$ ergibt sich nun als $\hat{\sigma}(Q) = S/n$.

Sei im folgenden der bivariate Fall betrachtet. Dazu sei $DR(h) = [l, r] - [l + h, r - h]$ der (innere) Randbereich der Domäne $D = [l, r]$ mit $l = (l_1, l_2)^t$, $r = (r_1, r_2)^t$ und $h = (h_1, h_2)^t$ sowie $QR(h) = [a - h, b + h]$ der (innere und äußere) Randbereich der Anfrage $Q(a,b)$ mit $a = (a_1,a_2)^t$ und $b = (b_1,b_2)^t$.

Durch Integration von Gleichung (2.74) ergibt sich im univariaten Fall eine analoge Formel wie in Gleichung (3.39). Diese ist jedoch für praktische Zwecke unbrauchbar. Stattdessen werden im Algorithmus zu Stichprobenwerten $\boldsymbol{X_i} = (X_{i1},X_{i2})^t$ aus $DR \cap AR$ nach folgender Fallunterscheidung die angegebenen gespiegelten Stichprobenwerte berücksichtigt:

(i) $a_1 < l_1 + h_1 : (2l_1 - X_{i1}, X_{i2})$

(ii) $b_1 > r_1 - h_1 : (2r_1 + X_{i1}, X_{i2})$

(iii) $a_2 < l_2 + h_2 : (X_{i1}, 2l_2 - X_{i2})$

(iv) $b_2 > r_2 - h_2 : (X_{i1}, 2r_2 + X_{i2}))$

(v) $a_1 < l_1 + h_1 \wedge a_2 < l_2 + h_2 : (2l_1 - X_{i1}, 2l_2 - X_{i2})$

(vi) $a_1 < l_1 + h_1 \wedge b_2 > r_2 - h_2 : (2l_1 - X_{i1}, 2r_2 + X_{i2})$

(vii) $b_1 > r_1 - h_1 \wedge a_2 < l_2 + h_2 : (2r_1 + X_{i1}, 2l_2 - X_{i2})$

(viii) $b_1 > r_1 - h_1 \wedge b_2 > r_2 - h_2 : (2r_1 + X_{i1}, 2r_2 + X_{i2})$

Somit ergibt sich der folgende Algorithmus:

Algorithmus 3.7 (*bivariate Kernselektivitätsschätzung mit Spiegelung*):

Gegeben: Dimension d=2, eine Relation R mit Instanz I und Stichprobe $X_1,...,X_n$ der Größe n, Bereichsanfrage $Q(a,b)$ mit $a = (a_1, a_2)$ und $b = (b_1, b_2)$ auf dem Intervall $[(l_1, l_2), (r_1, r_2)]$, Epanechnikow-Produktkernfunktion K mit Bandbreiten h_1 und h_2.

1. Setze S gleich der Anzahl der Stichprobenelemente X_i im Intervall $[(a_1 + h_1, a_2 + h_2), (b_1 - h_1, b_2 - h_2)]$.
2. Berechne für alle Stichprobenelemente X_i aus $[(a_1 - h_1, a_2 - h_2), (b_1 + h_1, b_2 + h_2)]$ das entsprechende Integral aus Gleichung (3.35) und addiere den Wert zu S hinzu.
3. Für die obigen Fälle (i) bis (viii) berechne das Integral aus Gleichung (3.35) der entsprechenden gespiegelten Stichprobenelemente und addiere die Werte zu S hinzu. Beachte, daß die obigen Fälle nicht ausschließlich sind. So können z.B. die Fälle (i), (iii) und (v) gleichzeitig auftreten.

3. Die geschätzte Selektivität der Anfrage $Q(a,b)$ ergibt sich nun als $\hat{\sigma}(Q) = S/n$.

Konsistenter Randkern mit niedrigem Bias

In Kapitel 2.3.5 wurden konsistente Randkerne mit niedrigem Bias ($O(h^2)$) beschrieben. Als für die Praxis besonders geeignet stellte sich dabei der Epanechnikow-Randkern von [Dong & Simonoff 94] dar (s. Gleichung (2.79)), der hier zur Selektivitätsschätzung angewendet wird.

Sei zunächst der univariate Fall mit Domäne $[l, r]$ und Anfrage $Q(a,b)$ behandelt. Sei h die gewählte Bandbreite. Sei weiterhin zunächst nur der linke Rand betrachtet mit $q = (x - l)/h$. Dazu ist nach obiger Fallunterscheidung u.a. das Integral $\int_a^\beta \hat{f}(t)dt$ mit $b = \min\{h,b\}$ unter Verwendung des linken Randkerns von [Dong & Simonoff 94] zu bestimmen. Dieser wird im folgenden kurz entwickelt:

$$\int_a^\beta \hat{f}(t)dt = \int_a^\beta \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K_{DS}^{l}\left(\frac{t - X_i}{h}, q(t)\right)dt = \frac{1}{n}\sum_{i=1}^{n} \int_a^\beta \frac{1}{h} K_{DS}^{l}\left(\frac{t - X_i}{h}, q(t)\right)dt \qquad (3.40)$$

Zur Berechnung des Integrales betrachte

$$\int_a^\beta \frac{1}{h} K_{DS}^{l}\left(\frac{t - X_i}{h}, q(t)\right)dt = \int_a^\beta \frac{1}{h} K_{DS}^{l}\left(\frac{t - X_i}{h}, q(t)\right)dt$$

$$= 3\int_a^\beta \frac{1}{h} \frac{1 + q^2 - 2\left(\frac{t - X_i}{h}\right)^2}{(1+q)^3} 1_{[-1,\,q]}dt$$

$$= 3\int_{(a-l)/h}^{(\beta-l)/h} \frac{1 + q^2 - 2\left(q - \frac{X_i - l}{h}\right)^2}{(1+q)^3} 1_{[-1,\,q]}dq \text{ (mit Transformation } t = qh+l)$$

$$= 3\left(-\ln(1+x) - 2\frac{1+\frac{X_i - l}{h}}{(1+x)} + \frac{(X_i - l)^2}{h^2} \right)\Bigg|_{(a-l)/h}^{(\beta - l)/h} \tag{3.41}$$

$$= F_{DS}^{l}(x;X_i, h)\Big|_{(a-l)/h}^{(\beta - l)/h}$$

In den Spezialfällen $b = l+h$ bzw. $a = l$ ergibt sich für die Funktion $F_{DS}^{l}(x;X_i, h)$

$$F_{DS}^{l}(l+h;X_i, h) = 3\left(-\ln(2) - \left(1 + \frac{X_i - l}{h}\right) + \frac{(X_i - l)^2}{h^2} \right) \text{ bzw.}$$

$$F_{DS}^{l}(l;X_i, h) = 3\left(-2\left(1 + \frac{X_i - l}{h}\right) + \frac{(X_i - l)^2}{h^2} \right).$$

Die Bestimmung des Randkerns von [Dong & Simonoff 94] am rechten Rand r mit $q = (r - x)/h$ erfolgt analog:

$$\int_{\alpha}^{b} \frac{1}{h} K_{DS}^{r}\left(\frac{t - X_i}{h}, q(t)\right) dt = \int_{\alpha}^{b} \frac{1}{h} K_{DS}^{r}\left(\frac{t - X_i}{h}, q(t)\right) dt$$

$$= 3\int_{\alpha}^{b} \frac{1}{h} \frac{1 + q^2 - 2\left(\frac{t - X_i}{h}\right)^2}{(1+q)^3} 1_{[-q,\,1]}\, dt$$

$$= -3 \int_{(r-\alpha)/h}^{(r-b)/h} \frac{1 + q^2 + 2\left(-q + \frac{r - X_i}{h}\right)^2}{(1+q)^3} 1_{[-q,\,1]}\, dq \text{ (mit Transformation } t = r - qh\text{)}$$

$$= 3\left(-\ln(1+x) - 2\frac{1+\frac{r - X_i}{h}}{(1+x)} + \frac{(r - X_i)^2}{h^2} \right)\Bigg|_{(r-\alpha)/h}^{(r-b)/h} \tag{3.42}$$

$$= F_{DS}^{r}(x;X_i, h)\Big|_{(r-\alpha)/h}^{(r-b)/h}$$

In den Spezialfällen $a = r - h$ bzw. $b = r$ ergibt sich für die Funktion $F_{DS}^{r}(x;X_i, h)$

$$F^r_{DS}(r-h;X_i,h) = 3\left(-\ln(2)-\left(1+\frac{r-X_i}{h}\right)+\frac{(r-X_i)^2}{h^2}\right) \text{ bzw.}$$

$$F^r_{DS}(r;X_i,h) = 3\left(-2\left(1+\frac{r-X_i}{h}\right)+\frac{(r-X_i)^2}{h^2}\right).$$

Im multivariaten Fall sind die jeweiligen Produktkerne zu wählen, vgl. Ende von Kapitel 2.3.5 sowie Gleichung (3.30).

Algorithmus 3.8 (*Kernselektivitätsschätzung mit Randkernfunktionen*):

Gegeben: Dimension d, eine Relation R mit Instanz I und Stichprobe $X_1,...,X_n$ der Größe n, Bereichsanfrage $Q(a,b)$ auf dem Intervall $[l,r]$, Epanechnikow-Produktkern K mit Bandbreite $H = diag(h_1^2,h_2^2)$ sowie Epanechnikow-Randkern K_{DS} von [Dong & Simonoff 94].

1. Setze S gleich der Anzahl der Stichprobenelemente X_i aus $I(a,b)$.
2. Berechne für alle Stichprobenelemente X_i aus $R(a,b)$ die zugehörigen Integrale gemäß Gleichung (3.35) und (3.26), sowie für X_i aus $E(a,b)$ das komplette Integral gemäß Gleichung (3.31) und addiere die Werte zu S hinzu.
3. Die geschätzte Selektivität der Anfrage $Q(a,b)$ ergibt sich nun als $\hat{\sigma}(Q) = S/n$.

In Kapitel 5 zeigen die Experimente die Überlegenheit der Kernschätzer mit der Technik der Spiegelung oder Randkernen über einfachen Kernschätzern bei Verteilungen mit Masse am Rand.

3.5 Hybridselektivitätsschätzer

Eine wichtige Voraussetzung für die Kernschätzer ist die Glattheit der den Daten zugrunde liegenden Dichtefunktion. Diese Annahme ist für reale Daten i.a. nicht haltbar. Besitzen die Daten Sprungstellen, so tauchen dort ähnliche Effekte auf wie an den Rändern - die Masse links und rechts der Sprungstelle wird gleichmäßig auf beide Seiten verteilt. Idee des Hybridselektivitätsschätzers ist es nun, an solchen Sprungstellen künstliche Rändern zu erzeugen und die Dichte bzw. Selektivität links und rechts des künstlichen Randes mit geeigneten Randbetrachtungen zu schätzen. In der hier vorgeschlagenen Methode werden daher mittels geeigneter Kriterien Histogramme gebildet, die den Datenraum in k glatte Teilbereiche partitionieren. Innerhalb der Histogrammbins wird nun die Gleichverteilungsannahme des Histogrammschätzers fallen gelassen. Stattdessen werden Kernselektivitätsschätzer mit entsprechender Randbehandlung verwendet. Die Verwendung von Randkernschätzern ist hier besonders bedeutsam, da an jedem Rand eines Histogrammbins ein Randproblem entstehen kann. Insbesondere ist i.a. davon auszugehen, daß an den Rändern der Histogramm-Bins die Dichte ungleich Null ist. Für jedes His-

togramm-Bin C_j muß die jeweilige Bandbreite h_j separat berechnet werden. Geht die Anfrage über mehrere Histogrammbins, so müssen die Einzel-Ergebnisse am Ende kombiniert werden zur Gesamt-Selektivitätsschätzung.

Eine weitere Eigenschaft der hier vorgeschlagenen Methode der Selektivitätsschätzung ist, daß bei der Kernselektivitätsschätzung in den jeweiligen Histogrammbins auch unterschiedliche Bandbreiten eingesetzt werden können und i.a. auch eingesetzt werden in Abhängigkeit von der dem Histogramm-Bin zugrunde liegenden Teil-Stichprobe. Es handelt sich daher auch um ein spezielles Verfahren zur Bestimmung einer variablen Bandbreite, vgl hierzu Kapitel 4. Bei der Bildung der Histogrammbins ist zu gewährleisten, daß die in den einzelnen Bins vorliegenden Stichproben ausreichend groß sind.

Ungeklärt ist bisher die Wahl der Anzahl der Histogramm-Bins und das Partitionierkriterium. Bei den Experimenten mit univariaten Testdaten in Kapitel 5.5.6 wird ein heuristisches Verfahren verwendet, daß sich allerdings nicht für praktische Implementierungen in DBMS eignet. Hier wird visuell anhand der Fehler der einzelnen Anfragen beurteilt, an welchen Stellen mögliche Sprungstellen vorliegen. Dabei wird davon ausgegangen, daß an Sprungstellen markante Extrema in der Fehlerkurve vorliegen. Die Ergebnisse der Experimente bestätigen die vorgeschlagene Vorgehensweise. Die visuelle Beurteilung eignet sich jedoch nicht zur praktischen Nutzung in DBMS, wo die Selektivitätsschätzung automatisch erfolgen muß. Hier besteht weiterer Forschungsbedarf. Ein möglicher Ansatz besteht zum Beispiel darin, Sprungstellen mit Hilfe von links- und rechtsseitigen Kernschätzern ([Qiu & Yandell 94], [Qiu 97]) zu detektieren.

Die Verwendung der klassischen Partitionierungsverfahren für Histogramme wie beim Equi-Width oder Equi-Depth Histogramm macht an dieser Stelle i.a. keinen Sinn, da die Grenzen der Histogramm-Bins i.a. nicht mit den vorhandenen Unstetigkeitsstellen übereinstimmen würden.

Sind für eine gegebene Anfrage $Q(a,b)$ die Einzelselektivitäten $\hat{\sigma}_j$ geschätzt, so ergibt sich die Schätzung der Gesamtselektivität $\hat{\sigma}_Y(Q)$ als

$$\hat{\sigma}_Y(Q) = \frac{1}{n}\sum_{j=1}^{k} n_j \hat{\sigma}_j(Q) \tag{3.43}$$

mit n_j der Anzahl der in C_j enthaltenen Stichprobenelemente.

Zusammengefaßt ergibt sich somit folgender Algorithmus für den univariaten Hybridschätzer:

Algorithmus 3.9 (*Hybridselektivitätsschätzer*):

Gegeben: Dimension d=1, eine Relation R mit Instanz I und Stichprobe $X_1,...,X_n$ der Größe n, Bereichsanfrage $Q(a,b)$ auf dem Intervall $[l,r]$, Epanechnikow-Kernfunktion K mit Bandbreite h.

1. Bestimme die Anzahl $k-1$ und die Lage möglicher Sprungstellen und bilde die zugehörigen Histogramm-Bins C_j, $j = 0,...,k-1$.
2. Für jedes C_j bestimme die Größe n_j der in C_j enthaltenen Teil-Stichprobe und berechne die geschätzte Teilselektivität $\hat{\sigma}_j$ mittels Kernselektivitätsschätzung mit Randbehandlung.
3. Die geschätzte Gesamtselektivität $\hat{\sigma}_Y(Q)$ der Anfrage $Q(a,b)$ ergibt sich nun gemäß Gleichung (3.43).

I.a. kommt es in den verschiedenen Bins zu einer unterschiedlichen Wahl der Bandbreite für den lokalen Kernselektivitätsschätzer. Es handelt sich bei diesem Schätzer also um ein Verfahren mit variabler Bandbreite, vgl. hierzu Kapitel 4.4.4. Hauptvorteil des Hybridschätzers liegt jedoch weiterhin in der Berücksichtigung von Sprungstellen.

Die Ergebnisse der Hybrid-Methode werden in Kapitel 5.5 präsentiert. Sie zeigen eine deutliche Verbesserung bei realen Datensätzen. Ein künstlicher Datensatz mit künstlichen Sprungstellen zeigt ebenfalls bei den Experimenten bessere Ergebnisse bei Verwendung der Hybridmethode.

3.6 Speicherbedarf

Zur effizienten Verwendung der Selektivitätsschätzer in Datenbankmanagementsystemen ist die Haltung der für die Schätzung relevanten Informationen über die Daten im Hauptspeicher des Rechners erforderlich. Deshalb sei im folgenden der Speicheraufwand für verschiedene Selektivitätsschätzer angegeben.

Sei dazu d die betrachtete Dimension, n die Stichprobengröße, k die Anzahl der Histogrammbins, m die Anzahl der Shifts beim Average Shifted Histogramm und b die Anzahl der Bytes, die zur Speicherung einer Realzahl notwendig sind. Es wird dabei im folgenden davon ausgegangen, daß alle notwendigen Informationen in Form von Realzahlen gespeichert werden.

Die Speicherung der Domänengrenze erfordert für alle Schätzer $2 \cdot d \cdot b$ Bytes und ist vernachlässigbar.

Der direkte Selektivitätsschätzer und der Kernselektivitätsschätzer benötigen lediglich eine Stichprobe der Größe n zur Berechnung der Selektivitätsschätzungen. Abhängig von der Dimension sind zur Speicherung einer Stichprobe $d \cdot n \cdot b$ Bytes erforderlich. Eine weitere Reduktion des Speicheraufwandes läßt sich hierbei durch geeignete Kompressionsverfahren z.B. bei Verwendung von Wavelets oder Interpolationsverfahren erzielen.

Zur Speicherung eines Histogrammes mit k Bins sind lediglich die Speicherung der $2k$ Bingrenzen sowie der k Selektivitäten der Stichprobe für die einzelnen Bins erforderlich. Dies macht einen Speicheraufwand in Abhängigkeit von der Dimension von $(2 \cdot d + 1) \cdot k \cdot b$ Bytes.

Bei einem Average Shifted Histogramm werden m Histogramme mit je k Bins einschließlich deren Selektivitäten auf der Stichprobe gespeichert. Dies erfordert die Speicherung von $(2 \cdot d + 1) \cdot k \cdot m \cdot b$ Bytes.

Den geringsten Speicheraufwand erfordert sicherlich der Selektivitätsschätzer auf Basis der Gleichverteilungsannahme. Hier muß lediglich der Mittelwert der Verteilung der Daten (b Bytes) gespeichert werden.

Die Schätzung der Bandbreite erfordert lediglich die Speicherung weniger statistischer Kenngrößen wie Standardabweichung, Interquartilsabstand oder Korrelationskoeffizient - vgl. Kapitel 4. Der dadurch notwendige Speicherbedarf ist von daher vernachlässigbar.

Der Speicherbedarf war in früheren Veröffentlichungen über Selektivitätsschätzer ein häufig diskutiertes Thema. So benötigt z.B. ein Histogramm mit 20 Bins sicherlich weniger Hauptspeicher (im univariaten Fall etwa 240 Byte bei 4 Byte pro Real-Zahl) als eine Stichprobe der Größe 2000 (im univariaten Fall etwa 8000 Byte bei 4 Byte pro Real-Zahl). Obwohl diese Diskussion teilweise in neueren Veröffentlichungen noch fortgesetzt wird ([Poosala & Ioannidis 97]), ist sie historisch zu sehen. In heutigen Zeiten, in denen von Terabyte-Datenbanken die Rede ist und Hauptspeicher im GByte-Bereich bei großen Datenbanksystemen durchaus üblich sind, spielt es eine untergeordnete Rolle, ob eine Verfahren einige kByte mehr oder weniger verbraucht.

4. Bestimmung des Bandbreitenparameters

Experimente mit Histogrammselektivitätsschätzern aber auch Kernselektivitätsschätzern haben gezeigt, daß es abhängig von der Stichprobengröße eine optimale Anzahl von Bins bzw. eine optimale Bandbreite gibt. Bei gleichverteilten Daten entspricht die optimale Anzahl von Bins bei Histogrammen genau einem Bin, was der Annahme einer Gleichverteilung über der Domäne entspricht. Bei anderen Verteilungen ergibt sich ein endlicher Wert größer als einem Bin. Bei wachsender Anzahl der Bins steigt der Fehler ab der optimalen Binweite wieder an. Abbildung 4.1 zeigt die Abhängigkeit des durchschnittlichen relativen Fehlers von der Anzahl der Bins bei einer Histogrammselektivitätsschätzung mit gleicher Binweite. Man beachte, daß die Skalierung der *x*-Achse nicht linear ist. Der Schätzung lag ein normalverteilter Datensatz mit 100.000 Werten zugrunde, die Anfragegröße der 1.000 Anfragen betrug 1% des Datenraums und die Schätzung beruhte auf einer Stichprobe von 2.000 Elementen. Die gestrichelte Kurve zeigt den durchschnittlichen relativen Fehler, wogegen die durchgezogene Linie den Fehler unter Verwendung der empirischen Verteilungsfunktion angibt. Es stellt sich somit die Frage, ob und wie sich die optimale Bandbreite mathematisch bestimmen läßt.

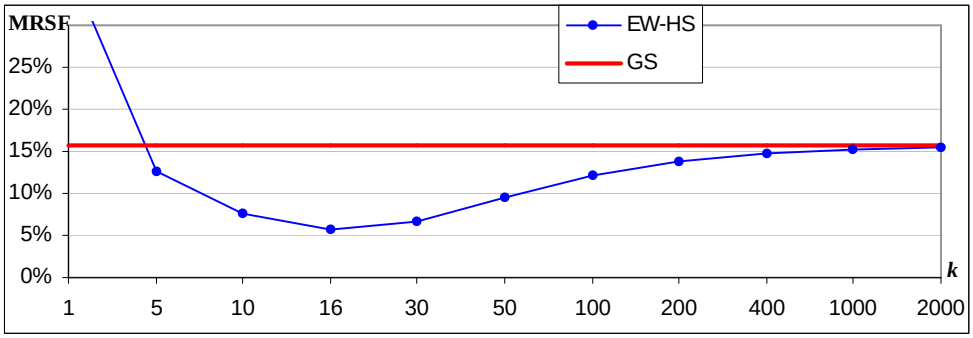


Abbildung 4.1: Abhängigkeit des MRSF von der Anzahl *k* der Bins beim Histogramm-Selektivitätsschätzer.

Die in den Experimenten beobachteten Verhaltensweisen des Schätzers in Abhängigkeit vom Bandbreitenparameter lassen sich mathematisch erklären. Dazu wird im folgenden der in Kapitel 2.2 eingeführte AMISE betrachtet und für verschiedene Schätzer angegeben (Abschnitt 4.1). Durch Minimierung des AMISE läßt sich ein optimaler Bandbreitenparameter bestimmen, der aber i.a. noch von der wahren Dichte *f* abhängt (Abschnitt 4.3). Ziel des vorliegenden Ansatzes ist nun die Bestimmung einer Bandbreite, die den AMISE minimiert. Diese sei im folgenden die *asymptotisch optimale Bandbreite AOB* genannt und mit h_{AO} im univariaten Fall und mit H_{AO} im multivariaten Fall bezeichnet. Die asymptotisch optimale Anzahl von Histogramm-Bins kann aus der AOB im univariaten Fall mittels Division des Wertebereichs durch h_{ao} bestimmt werden. Man beachte, daß die AOB i.a. nun immer noch von der unbekannten wahren Dichte *f* abhängt. Unter gewissen Annahmen an die wahre Dichte *f* können nun verschiedene Verfahren zur Schätzung des asymptotisch optimalen Bandbreitenparameter bei gegebener Stichprobengröße für verschiedene Schätzfunktionen angegeben werden

(Abschnitt 4.4). Dabei werden auch Verfahren vorgestellt, die nicht von einer festen AOB für die gesamte Domäne ausgehen, sondern eine variable AOB angeben in Abhängigkeit von der lokalen Dichte. Bei gegebenem AMISE und asymptotisch optimalem Bandbreitenparameter läßt sich als ein weiteres Gütekriterium die Konvergenzordnung eines Schätzers angeben (Abschnitt 4.5). Umgekehrt läßt sich bei vorgegebener Fehlerschranke des AMISE die zur Einhaltung notwendige Stichprobengröße angeben (Abschnitt 4.7).

Bei der Bestimmung der Bandbreite wird im folgenden i.a. davon ausgegangen, daß ein reguläres Gitter vorliegt, d.h. die Bandbreite ist konstant für jede Dimension. [Scott 92] hat im bivariaten Fall quadratische, dreieckige und hexagonale Gitterstrukturen auf Basis des AMISE für Histogrammschätzer miteinander verglichen. Hexagonale Gittestrukturen liefern hier einen leicht niedrigeren AMISE als quadratische Gitterstrukturen, während dreieckige Gitterstrukturen etwas schlechter sind. Allerdings ist der Unterschied marginal, so daß sich der Mehraufwand bei hexagonalen Gitterstrukturen für praktische Anwendungen nicht lohnt. In Abschnitt 4.4.4 wird auf Schätzer mit variabler Bandbreite eingegangen, die hingegen durchaus bessere Ergebnisse liefern.

Als nützliche abkürzende Schreibweise diene die folgende Definition für jede reelle quadratisch-integrierbare Funktion g:

$$R(g) = \int g(x)^2 dx \tag{4.1}$$

Im folgenden wird der oben beschriebene generelle Ansatz auf verschiedene Schätzfunktionen angewendet. Die Herleitung der Gleichungen findet sich z.B. in [Scott 92], [Silverman 86] oder [Wand & Jones 95].

Verschiedene der vorgestellten Verfahren zur Bestimmung eines optimalen Bandbreitenparameter werden insbesondere bei Kernselektivitätsschätzern experimentell untersucht - die Ergebnisse sind in Kapitel 5 präsentiert.

4.1 Asymptotischer MISE für verschiedene Schätzfunktionen

Ausgangspunkt ist im folgenden der Bias und die Varianz der verwendeten Schätzfunktion $\hat{f}$. Hieraus läßt sich durch asymptotische Betrachtungen und Integration der AMISE herleiten. Der AMISE ist für die in Kapitel 2.3 vorgestellten Schätzer angegeben. Für die in dieser Arbeit am genauesten untersuchten Schätzer, den Equi-Width-Histogrammschätzer und den Kernschätzer, ist die Herleitung ausführlicher beschrieben.

Ein wichtiges Hilfsmittel ist hierbei die Anwendung der Taylor-Entwicklung einer (k+1)-mal differenzierbaren reellen Funktion $f(y)$ um einen Entwicklungspunkt x, vgl. [Forster 79]:

$$f(y) = \sum_{i=0}^{k} \frac{1}{i!} f^{(i)}(x)(y-x)^i + \frac{1}{k!} \int_x^y (y-t)^k f^{(k+1)}(t)dt \tag{4.2}$$

AMISE für Histogrammschätzer

Sei zunächst der univariate Fall betrachtet. Im folgenden sei dazu vorausgesetzt, daß die wahre Dichte f mindestens zweimal stetig differenzierbar und Lipschitz-stetig über jedem Histogramm-Bin C_j, $j = 1...k$, ist.

Für den Erwartungswert und die Varianz des univariaten Histogrammschätzers mit fester Bandbreite h waren bereits in Kapitel 2.3.2 die Gleichungen (2.35) und (2.36) angegeben. Daraus ergibt sich weiterhin für den Bias:

$$Bias(\hat{f}_H(x)) = \frac{p_j}{h} - f(x) \text{ mit}$$

$$p_j = \int_{jh}^{(j+1)h} f(t)dt = \int_{jh}^{(j+1)h} \left(f(x) + (t-x)f^{(1)}(x) + \frac{1}{2}(t-x)^2 f^{(2)}(x) + \ldots \right) dt$$

$$= h\left(f(x) + \left(\frac{h}{2} - x \right) f^{(1)}(x) \right) + O(h^3) \quad \Rightarrow$$

$$Bias(\hat{f}_H(x)) = \left(\frac{h}{2} - x \right) f^{(1)}(x) + O(h^2) \tag{4.3}$$

und für die Varianz:

$$Var(\hat{f}_H(x)) = \frac{1}{nh^2}(p_j - p_j^2) \text{ für } x \in C_j$$

$$\text{mit } p_j = \int_{C_j} f(t)dt = hf(\xi_j) \text{ für ein } \xi_j \in C_j$$

nach dem Mittelwertsatz der Integralrechnung [Forster 79] (4.4)

Sodann ergibt sich aus Gleichung (4.3) und aus

$$\int_{C_0} \left(\frac{h}{2} - x \right)^2 (f^{(1)}(x))^2 = \int_0^h \left(\frac{h}{2} - x \right)^2 (f^{(1)}(x))^2 dx$$

$= (f^{(1)}(\zeta_j))^2 \int_0^h \left(\frac{h}{2} - x\right)^2 dx$ für ein $\zeta_j \in C_j$ nach dem verallgemeinerten Mittelwertsatz der Integralrechnung,

$= \frac{h^3}{12}(f^{(1)}(\zeta_j))^2$

für den integrierten quadratischen Bias (IQBias):

$$IQBias(\hat{f}_H(x)) = \frac{h^2}{12}\int (f^{(1)}(x))^2 dx + o(h^2) = \frac{h^2}{12}R(f^{(1)}) + o(h^2) \tag{4.5}$$

Für die integrierte Varianz (IVar) ergibt sich aus Gleichung (4.4):

$$IVar(\hat{f}_H(x)) = \int Var(\hat{f}_H(x))dx = \sum_j \int_{Cj} Var(\hat{f}_H(x))dx$$

$$= \sum_j \int_{Cj} \frac{1}{nh^2}(p_j - p_j^2)dx = \frac{1}{nh}\sum_j (p_j - p_j^2)$$

$= \frac{1}{nh}\int f(x)dx - \frac{1}{nh}\sum_j (f(\xi_j)h)^2$ für jeweils ein $\zeta_j \in C_j$ nach dem Mittelwertsatz der Integralrechnung

$= \frac{1}{nh} - \frac{1}{nh}h(\int (f(x))^2 dx + o(1)) = \frac{1}{nh} - \frac{1}{n}\int (f(x))^2 dx + o\left(\frac{1}{n}\right)$ mit $\int f(x)dx = 1$

$$= \frac{1}{nh} - \frac{R(f)}{n} + o\left(\frac{1}{n}\right) \tag{4.6}$$

Insgesamt ergibt sich somit für den asymptotischen MISE des univariaten Equi-Width-Histogrammschätzers $\hat{f}_H$ (bei n unabhängigen gleichverteilten Stichprobenelementen und in Abhängigkeit von der Bandbreite h):

$$AMISE(\hat{f}_H(x)) = \frac{1}{nh} + \frac{1}{12}h^2 R(f') \tag{4.7}$$

Die Rechnungen im bivariaten bzw. multivariaten Fall erfolgen analog. Im bivariaten Fall ist die Varianz des Equi-Width-Histogrammschätzers gegeben durch

$$Var(\hat{f}_H(\boldsymbol{x})) = \frac{p_j(1-p_j)}{n(h_1 h_2)^2}$$

und der Bias durch

$$Bias(\hat{f}_H(\boldsymbol{x})) = \frac{p_j}{h_1 h_2} - f(\boldsymbol{x})$$

Daraus folgt für die integrierte Varianz

$$IVar(\hat{f}_H(\boldsymbol{x})) = \frac{1}{n h_1 h_2} - \frac{R(f)}{n} + o\left(\frac{1}{n}\right)$$

und den integrierten quadratischen Bias

$$IQBias(\hat{f}_H(\boldsymbol{x})) = \frac{{h_1}^2}{12} R\left(\frac{\mathrm{d}}{\mathrm{d}x_1} f(\boldsymbol{x})\right) + \frac{{h_2}^2}{12} R\left(\frac{\mathrm{d}}{\mathrm{d}x_2} f(\boldsymbol{x})\right) + o(h^4)$$

Zusammen ergibt sich daraus der AMISE des bivariaten Equi-Width-Histogrammschätzers

$$AMISE(\hat{f}_H(\boldsymbol{x})) = \frac{1}{n h_1 h_2} + \frac{1}{12}(h_1^2 R(f_1) + h_2^2 R(f_2)) \text{ mit } f_i = \frac{d}{dx_i} f(\boldsymbol{x}) \tag{4.8}$$

Analog ergibt sich für den multivariaten Equi-Width-Histogrammschätzer

$$AMISE(\hat{f}_H(\boldsymbol{x})) = \frac{1}{n}\left(\prod_{j=1}^{d} h_j\right)^{-1} + \frac{1}{12}\sum_{j=1}^{d} h_j^2 R(f_j) \text{ mit } f_i = \frac{d}{dx_j} f(\boldsymbol{x}),\ j = 1...d. \tag{4.9}$$

AMISE für Häufigkeitspolygonschätzer

Der im folgenden angegebene univariate und multivariate AMISE für den Häufigkeitspolygonschätzer ist [Scott 92] entnommen. Dabei wird vorausgesetzt, daß die erste Ableitung der wahren Dichte *f* stetig und quadratisch integrierbar ist.

Der univariate AMISE für den Häufigkeitspolygonschätzer ist gegeben durch ([Scott 92], S. 98):

$$AMISE(\hat{f}_{FP}(x)) = \frac{2}{3nh} + \frac{49}{2880} h^4 R(f^{(2)}) \tag{4.10}$$

Der multivariate AMISE für den Häufigkeitspolygonschätzer ist gegeben durch ([Scott 92], S. 107):

$$AMISE(\hat{f}_{FP}(\boldsymbol{x})) = \left(\frac{2}{3}\right)^d \frac{1}{n}\left(\prod_{j=1}^{d} h_i\right)^{-1} + \frac{49}{2880}\sum_{j=1}^{d} h_i^4 R(f_{jj})$$

$$\text{mit } f_{jj} = \frac{\mathrm{d}^2}{\mathrm{d}x_j} f(\boldsymbol{x}),\ j = 1...d. \tag{4.11}$$

AMISE für Average Shifted Histogrammschätzer

Der im folgenden angegebene univariate und multivariate AMISE für den Average Shifted Histogrammschätzer ist ebenfalls [Scott 92] entnommen.

Der univariate AMISE für den Average Shifted Histogrammschätzer ist gegeben durch ([Scott 92], S. 119):

$$AMISE(\hat{f}_{ASH}(x)) = \frac{2}{3nh}\left(1 + \frac{1}{2m^2}\right) + \frac{h^2}{12m^2} R(f^{(1)}) + \frac{h^4}{144} R(f^{(2)})\left(1 - \frac{2}{m^2} + \frac{3}{5m^4}\right) \tag{4.12}$$

Bemerkung 4.1:

> Wählt man $m = 1$, so ergibt sich der AMISE für den univariaten Histogrammschätzer, vgl. Gleichung (4.7).

Der multivariate AMISE für den Average Shifted Histogrammschätzer ist gegeben durch ([Scott 92], S. 120):

$$AMISE(\hat{f}_{ASH}(\boldsymbol{x})) = \left(\frac{2}{3}\right)^d \frac{1}{n}\left(\prod_{j=1}^{d} h_j\right)^{-1} + \frac{1}{720}\sum_{j=1}^{d} \delta_j^4 R(f_{jj}) + \frac{1}{144}\int_{\Re^d}\left(\sum_{j=1}^{d} h_j^2\left(1 + \frac{1}{2m_j^2}\right) f_{jj}\right)^2 d\boldsymbol{x} \tag{4.13}$$

$$\text{mit } f_{jj} = \frac{\mathrm{d}^2}{\mathrm{d}x_j} f(\boldsymbol{x}),\ j = 1...d.$$

AMISE für Kernschätzer

Im Falle des Kernschätzers werden gewisse Voraussetzungen an die Kernfunktion und die wahre Dichte gemacht, damit die Konvergenz des AMISE sichergestellt ist. Hierzu existieren

verschiedene Möglichkeiten. Die folgenden Annahmen sind aus [Scott 92] übernommen. Andere Annahmen finden sich z.B. in [Wand & Jones 95], S.19-20.

Dazu habe die wahre Dichte f stetige Ableitungen aller erforderlichen Ordnungen und die zweite Ableitung sei quadratisch integrierbar und monoton.

Für die Bandbreite $h = h_n$, $h_n > 0$ gelte weiterhin $\lim_{n \to \infty} h = 0$ und $\lim_{n \to \infty} nh = \infty$. Damit wird sichergestellt, daß die Bandbreite bei wachsender Stichprobengröße gegen 0 strebt, aber nicht so stark wie die Stichprobengröße gegen unendlich.

Des weiteren sei im folgenden vorausgesetzt, daß K eine nicht-negative, symmetrische und stetige Kernfunktion 2. Ordnung sei, so daß für die Momente im univariaten Fall gilt:

$$\begin{aligned} &\text{(a) } \kappa_0(K) = 1 \\ &\text{(b) } \kappa_1(K) = 0 \\ &\text{(c) } \kappa_2(K) \neq 0 \end{aligned} \tag{4.14}$$

bzw. im multivariaten Fall:

$$\begin{aligned} &\text{a) } \kappa_0(K) = \int_{\Re^d} K(\boldsymbol{t})d\boldsymbol{t} = \mathbf{1}, \\ &\text{b) } \kappa_1(K) = \int_{\Re^d} \boldsymbol{t}K(\boldsymbol{t})d\boldsymbol{t} = \mathbf{0} \text{ und} \\ &\text{c) } \kappa_2(K) = \int_{\Re^d} \boldsymbol{t}\boldsymbol{t}^T K(\boldsymbol{t})d\boldsymbol{t} = c\boldsymbol{I}_d \end{aligned} \tag{4.15}$$

Bemerkung 4.2:

Folgende Kernfunktionen erfüllen beispielsweise diese Forderungen mit entsprechendem $k_2(K)$

- im univariaten Fall:

Kernfunktion	$k_2(K)$
Epanechnikow-Kern	1/5
Biweight-Kern	1/7
Gauß-Kern	1

Tabelle 4.1: Voraussetzung $k_2(K) > 0$ bei speziellen Kernfunktionen im univariaten Fall

- im bivariaten Fall beim Produktkern:

Kernfunktion	$k_2(K)$
Epanechnikow-Kern	2/5
Biweight-Kern	2/7
Gauß-Kern	2

Tabelle 4.2: Voraussetzung $k_2(K) > 0$ bei speziellen Kernfunktionen im bivariaten Fall

Bereits in Kapitel 2.3.4 wurden Gleichungen für den Erwartungswert und die Varianz des Kernschätzers angegeben (Gleichung (2.58)). Diese werden im folgenden verwendet um eine Gleichung für den AMISE herzuleiten. Für den Bias des Kernschätzers gilt zunächst:

$$
\begin{aligned}
Bias_h(\hat{f}_K(x)) &= E(K_h(x, X)) - f(x) \\
&= \frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)dy - f(x) = \int K(t)f(x-ht)dt - f(x) \text{ mit } t = \frac{x-y}{h}, \\
&= \int K(t)\left(f(x) - f^{(1)}(x)th + \frac{1}{2}f^{(2)}(x)(th)^2 + \ldots\right)dt - f(x) \text{ mit (4.2)}, \\
&= f(x)\int K(t)dt - f^{(1)}(x)h\int tK(t)dt + \frac{h^2}{2}f^{(2)}(x)(\int t^2K(t)dt) + \ldots - f(x) \\
&= (f(x)\cdot 1) - (f^{(1)}(x)h \cdot 0) + \frac{h^2}{2}f^{(2)}(x)\kappa_2(K) + O(h^3) - f(x) \text{ mit (4.14)}, \\
&= \frac{h^2}{2}f^{(2)}(x)\kappa_2(K) + O(h^3) \qquad (4.16)
\end{aligned}
$$

Ebenso gilt für die Varianz:

$$
\begin{aligned}
Var_h(\hat{f}_K(x)) &= \frac{1}{n}E(\hat{f}^2(x)) - E(\hat{f}(x))^2 \\
&= \frac{1}{n}(\int K_h(x-y)^2 f(y)dy - (\int K_h(x-y)f(y)dy)^2) \\
&= \frac{1}{n}(\int K(t)^2 f(x-th)dt - (\int K(t)f(x-th)dt)^2) \\
&= \frac{1}{n}(\int K(t)^2 f(x-th)dt - (Bias(\hat{f}(x)) + f(x))^2) \\
&= \frac{1}{n}\left(\frac{1}{h}\int K(t)^2 f(x-th)dt - (f(x) + O(h^2))^2\right)
\end{aligned}
$$

$$= \frac{1}{nh}\left(\int K(t)^2 (f(x) - htf^{(1)}(x) + \ldots)dt\right) - \frac{1}{n}(f^2(x) + f(x)O(h^2)) \text{ mit (4.2)},$$

$$= \frac{1}{nh}f(x)R(K) - \frac{1}{n}f^2(x) + O\left(\frac{h}{n}\right) \text{ mit (4.1)}. \qquad (4.17)$$

Durch Integration ergibt sich aus Gleichung (4.16) als integriertem quadrierten Bias (*IQBias*):

$$IQBias_h(\hat{f}_K(x)) = \int\left(\frac{h^2}{2}f^{(2)}(x)\kappa_2(K)\right)^2 dx + \ldots$$

$$= \frac{h^4}{4}\kappa_2^2(K)R(f^{(2)}) + \ldots \text{ mit (4.1)}. \qquad (4.18)$$

und aus Gleichung (4.17) als integrierte Varianz (*IVar*):

$$IVar(\hat{f}_K(x)) = \int\left(\frac{1}{nh}f(x)R(K) - \frac{1}{n}f^2(x)\right)dx + \ldots$$

$$= \frac{1}{nh}R(K)\int f(x)dx - \frac{1}{n}\int f(x)^2 dx + \ldots$$

$$= \frac{1}{nh}R(K) - \frac{1}{n}R(f) + \ldots \text{ mit (4.1)}. \qquad (4.19)$$

Somit ergibt sich der AMISE des univariaten Kernschätzers aus den Gleichungen (4.18) und (4.19):

$$AMISE(\hat{f}_K(x)) = \frac{1}{nh}R(K) + \frac{h^4}{4}\kappa_2^2(K)R(f^{(2)}) \qquad (4.20)$$

Sei im folgenden ein bivariater Produktkern zweiter Ordnung mit Diagonal-Bandbreiten-Matrix $H = \text{diag}(h_1,h_2)$ betrachtet. Dann läßt sich der bivariate Erwartungswert berechnen als:

$$E(\hat{f}_{K^P}(\boldsymbol{x})) = E(K_{h_1}(x_1 - X_{i1})K_{h_2}(x_2 - X_{i2}))$$

$$= \int_{\Re^2} K_{h_1}(x_1 - y_1)K_{h_2}(x_2 - y_2)f(\boldsymbol{y})d\boldsymbol{y}$$

$$= \int_{\Re^2} K(t_1)K(t_2)f(x_1 + t_1h_1, x_2 + t_2h_2)d\boldsymbol{t}$$

$$= \int_{\Re^2} K(\boldsymbol{t})\left(f(\boldsymbol{x}) - h_1 t_1 \frac{\mathrm{d}}{\mathrm{d}x_1} f(\boldsymbol{x}) - h_2 t_2 \frac{\mathrm{d}}{\mathrm{d}x_2} f(\boldsymbol{x}) + \frac{h_1^2}{2} t_1^2 \frac{\mathrm{d}^2}{\mathrm{d}x_1} f(\boldsymbol{x}) \right.$$
$$\left. + h_1 h_2 t_1 t_2 \frac{\mathrm{d}}{\mathrm{d}x_1}\frac{\mathrm{d}}{\mathrm{d}x_2} f(\boldsymbol{x}) + \frac{h_2^2}{2} t_2^2 \frac{\mathrm{d}^2}{\mathrm{d}x_2} f(\boldsymbol{x}) + \ldots \right) d\boldsymbol{t}$$

mit Taylorentwicklung von *f* um *b*,

$$= f(\boldsymbol{x}) + \frac{\kappa_2(K)}{2}\left(h_1^2 \frac{\mathrm{d}^2}{\mathrm{d}x_1} f(\boldsymbol{x}) + h_1 h_2 \frac{\mathrm{d}}{\mathrm{d}x_1}\frac{\mathrm{d}}{\mathrm{d}x_2} f(\boldsymbol{x}) + h_2^2 \frac{\mathrm{d}^2}{\mathrm{d}x_2} f(\boldsymbol{x}) \right) + O(h^4)$$

mit (4.15),

so daß sich für den bivariaten Bias ergibt:

$$Bias(\hat{f}_{K^P}(\boldsymbol{x})) \approx \frac{\kappa_2(K)}{2}\left(h_1^2 \frac{\mathrm{d}^2}{\mathrm{d}x_1} f(\boldsymbol{x}) + h_1 h_2 \frac{\mathrm{d}}{\mathrm{d}x_1}\frac{\mathrm{d}}{\mathrm{d}x_2} f(\boldsymbol{x}) + h_2^2 \frac{\mathrm{d}^2}{\mathrm{d}x_2} f(\boldsymbol{x}) \right) \tag{4.21}$$

Ebenso läßt sich für die bivariate Varianz zeigen:

$$Var(\hat{f}_{K^P}(\boldsymbol{x})) \approx \frac{1}{n h_1 h_2} f(\boldsymbol{x}) R(K)^2 - \frac{1}{n} f^2(\boldsymbol{x}) \tag{4.22}$$

Damit ergibt sich aus Gleichung (4.21) für den bivariaten integrierten quadrierten Bias:

$$IQBias(\hat{f}_{K^P}(\boldsymbol{x})) \approx \frac{\kappa_2^2(K)}{4}\left(h_1^4 R\left(\frac{\mathrm{d}^2}{\mathrm{d}x_1} f(\boldsymbol{x}) \right) + h_2^2 R\left(\frac{\mathrm{d}^2}{\mathrm{d}x_2} f(\boldsymbol{x}) \right) \right. \tag{4.23}$$
$$\left. + h_1 h_2 \int \frac{\mathrm{d}}{\mathrm{d}x_1}\frac{\mathrm{d}}{\mathrm{d}x_2} f(\boldsymbol{x}) d\boldsymbol{x} \right)$$

und aus Gleichung (4.22) für die bivariate integrierte Varianz:

$$IVar(\hat{f}_{K^P}(\boldsymbol{x})) \approx \frac{1}{n h_1 h_2} R(K)^2 - \frac{1}{n} R(f) \tag{4.24}$$

Zusammen ergibt sich aus (4.23) und (4.24) folgende Gleichung für den AMISE des bivariaten Produktkernschätzers mit Diagonalbandbreitenmatrix:

$$AMISE(\hat{f}_{K^P}(\boldsymbol{x})) = \frac{k_2^2(K)}{4}\left(h_1^4 R(f_{11}) + h_2^4 R(f_{22}) + 2h_1^2 h_2^2 \int_{\Re^2} f_{11} f_{22} dx\right) + \frac{R(K)^2}{nh_1 h_2} - \frac{R(f)}{n} \tag{4.25}$$

mit $f_{jj} = \frac{\mathrm{d}^2}{\mathrm{d}x_j} f(\boldsymbol{x})$, $j = 1...d$.

Ganz analog ergibt sich als AMISE für den multivariaten Kernschätzer bei einem Produktkern 2. Ordnung:

$$AMISE(\hat{f}_{K^P}(\boldsymbol{x})) = \frac{1}{4}\kappa_2^2(K)\left(\sum_{j=1}^{d} h_j^4 R(f_{jj}) + \sum_{i \neq j} h_i^2 h_j^2 \int_{\Re^d} f_{ii} f_{jj} dx\right) + \frac{(R(K))^d}{n}\left(\prod_{j=1}^{d} h_j\right)^{-1} - \frac{R(f)}{n} \tag{4.26}$$

Im folgenden sei noch $R(K)$ für einige bekannte Kernfunktionen angegeben.

- im univariaten Fall:

Kernfunktion	$\boldsymbol{R(K)}$
Epanechnikow-Kern	3/5
Biweight-Kern	5/7
Gauß-Kern	$1/(2\sqrt{\pi})$

Tabelle 4.3: Ausdruck $R(K)$ bei speziellen Kernfunktionen im univariaten Fall

- im multivariaten Fall beim Produktkern K^P mit $d > 1$:

Kernfunktion	$\boldsymbol{R(K^P)}$
Epanechnikow-Kern	$(3/5)^d$
Biweight-Kern	$(5/7)^d$
Gauß-Kern	$(2\sqrt{\pi})^{-d}$

Tabelle 4.4: Ausdruck $R(K)$ bei speziellen Produktkernen im multivariaten Fall

4.2 Einfluß der Bandbreite auf Bias und Varianz

Sowohl beim Histogrammschätzer als auch beim Kernschätzer sieht man leicht, daß sich deren AMISE in den asymptotisch integrierten Bias (AIBias) und die asymptotisch integrierte Varianz (AIVar) aufteilen lassen. Diese sind z.B. beim univariaten Equi-Width-Histogrammschätzer (vgl. Gleichung (4.7))

$$\text{(a)} AIBias = \frac{h^2}{12} R(f')$$

$$\text{(b)} AIVar = \frac{1}{nh}$$

und beim univariaten Kernschätzer (vgl. Gleichung (4.20))

$$\text{(a)} AIBias = \frac{1}{4} h^4 k_2^2 R(f'')$$

$$\text{(b)} AIVar = \frac{1}{nh} R(K)$$

Dabei fällt auf, daß der AIBias im Gegensatz zur AIVar unabhängig von *n* ist. Eine Erhöhung der Stichprobengröße führt somit zu einer geringeren AIVar, aber nicht zu einem geringeren oder höheren AIBias.

Vergleicht man weiter die beiden asymptotischen Teilausdrücke AIBias und AIVar miteinander, so wird ein fundamentales Problem der Dichteschätzung mit den vorgestellten Dichteschätzern deutlich. Verändern der Bandbreite zeigt ein widersprüchliches Verhalten bzgl. Bias und Varianz. Wird die Bandbreite vermindert um den Bias zu verkleinern, so führt dies zu einer größeren Varianz. Umgekehrt hat die Vergrößerung der Bandbreite zur Reduzierung der Varianz eine Erhöhung des Bias zur Folge. Ziel ist es daher eine Bandbreite zu finden, die den AMISE bzgl. beider Kriterien minimiert.

4.3 Asymptotisch optimale Bandbreite für verschiedene Schätzfunktionen

Ausgangspunkt der folgenden Betrachtungen ist der AMISE, der in Abschnitt 4.1 für die einzelnen betrachteten Schätzer angegeben ist. Hierbei wird der AMISE betrachtet als eine Funktion in Abhängigkeit von der (festen) Bandbreite *h*. Zur Berechnung des Minimums werden die Nullstellen der 1. Ableitung bestimmt. Hierzu wird der AMISE bzgl. *h* differenziert und der resultierende Ausdruck gleich 0 gesetzt. Anschließend kann die Gleichung nach *h* aufgelöst werden. Die Minima sind nun an den Nullstellen der 1. Ableitung, an denen die 2. Ableitung größer 0 ist. Diese Minima sind nun die gesuchten AOBs. Die resultierenden AOBs sind im folgenden für die verschiedenen Schätzfunktionen angegeben.

AOB für Histogrammschätzer

Differenzieren von Gleichung (4.7) bzgl. *h* ergibt:

$$\frac{d}{dh}(AMISE) = -\frac{1}{nh^2} + \frac{h}{6}R(f^{(1)})$$

Wird der Ausdruck gleich 0 gesetzt, so führt Auflösen nach der Bandbreite *h* zur univariaten AOB:

$$h_{EW\text{-}ao} = \left(\frac{6}{R(f^{(1)})}\right)^{1/3} \cdot n^{-1/3} \tag{4.27}$$

Hieraus kann die Anzahl der Histogramm-Bins leicht berechnet werden.

Im bivariaten Fall ergibt sich aus Gleichung (4.8) durch partielles Differenzieren:

$$\frac{d}{dh_1}(AMISE) = -\frac{1}{nh_1^2h_2} + \frac{1}{6}h_1R(f_1)$$

Auflösen nach h_2 ergibt (nach h_1 analog):

$$\frac{6}{nh_1^3R(f_1)} = h_2$$

Einsetzen und auflösen ergibt weiterhin:

$$h_2^8 = \frac{6^2R(f_1)}{n^2(R(f_2))^3} = (R(f_2))^{-4}6^2R(f_1)R(f_2)n^{-2}$$

Daraus folgt für das asymptotisch optimale h_2 (h_1 analog) des bivariaten Equi-Width-Histogrammschätzers:

$$h_{EW-ao,j} = (R(f_j))^{-\frac{1}{2}}6^{\frac{1}{4}}(R(f_1)R(f_2))^{\frac{1}{8}}n^{-\frac{1}{4}}, j = 1,2 \tag{4.28}$$

Die Berechnung im multivariaten Fall verlaufen analog. Differenzieren von Gleichung (4.9) nach h_1 (für h_2, ..., h_d analog) ergibt:

$$\frac{d}{dh_1}(AMISE) = -\frac{1}{nh_1^2 h_2 \ldots h_d} + \frac{1}{6}h_1 R(f_1)$$

Gegenseitiges Einsetzen und Auflösen ergibt für das asymptotisch optimale h_j des multivariaten Equi-Width-Histogrammschätzers, vgl. [Scott 92], S. 82:

$$h_{EW\text{-}ao,\,j} = (R(f_j))^{-\frac{1}{2}} 6^{\frac{1}{d+2}} \prod_{i=1}^{d} (R(f_i))^{\frac{1}{2(d+2)}} n^{-\frac{1}{d+2}}. \tag{4.29}$$

AOB für Häufigkeitspolygonschätzer

Differenzieren von Gleichung (4.10) bzgl. *h* ergibt:

$$\frac{d}{dh}(AMISE) = -\frac{1}{3nh^2} + \frac{49h^3}{720}R(f^{(2)})$$

Wird der Ausdruck gleich 0 gesetzt, so führt Auflösen nach der Bandbreite *h* zur AOB des univariaten Häufigkeitspolygonschätzers:

$$h_{FP\text{-}ao} = 2\left(\frac{15}{49R(f^{(2)})}\right)^{1/5} n^{-1/5} \tag{4.30}$$

Eine geschlossene Form für multivariate Häufigkeitspolygonschätzers läßt sich i.a. nicht angeben, vgl. [Scott 92], S. 107.

AOB für Average Shifted Histogrammschätzer

Für den univariaten Average Shifted Histogrammschätzer läßt sich ein einfacher Ausdruck für die asymptotisch optimale Bandbreite angeben, wenn man $m \to \infty$ betrachtet. Dann reduziert sich der AMISE zu:

$$AMISE(\hat{f}_{ASH,\,m\to\infty}(x)) = \frac{2}{3nh} + \frac{h^4}{144}R(f^{(2)})$$

Differenzieren des AMISE nach *h* ergibt:

$$\frac{d}{dh}(AMISE) = -\frac{2}{3nh^2} + \frac{h^3}{36}R(f^{(2)})$$

Nach h auflösen ergibt dann die AOB für den univariaten Average Shifted Histogrammschätzer mit $m \to \infty$:

$$h_{ASH,\,ao} = \left(\frac{24}{R(f^{(2)})} \right)^{1/5} n^{-1/5} \tag{4.31}$$

AOB für Kernschätzer

Sei zunächst der univariate Fall betrachtet. Differenzieren von Gleichung (4.20) nach h ergibt

$$\frac{d}{dh}(AMISE) = -\frac{1}{nh^2} R(K) + h^3 \kappa_2^2(K) R(f^{(2)}) \tag{4.32}$$

Setzen des Ausdrucks gleich 0 und Auflösen nach h ergibt als asymptotisch optimale Bandbreite für den Kernschätzer:

$$h_{K\text{-}ao} = \left(\frac{R(K)}{\kappa_2^2(K) R(f^{(2)})} \right)^{1/5} \cdot n^{-1/5} \tag{4.33}$$

Im multivariaten Fall existiert keine allgemeine geschlossene Form für die asymptotisch optimale Bandbreite, vgl. [Scott 92], S. 150f. Ergebnisse können lediglich für Spezialfälle angegeben werden, z.B. wenn $h_i = h$ für alle i, oder wenn für die wahre Dichte die Normalverteilung angenommen wird, siehe Abschnitt 4.4.1.

4.4 Verschiedene Verfahren zur Schätzung der asymptotisch optimalen Bandbreite

Die für die verschiedenen Schätzverfahren abgeleiteten AOB hängen alle von der noch unbekannten wahren Dichtefunktion f bzw. einer ihrer Ableitungen ab. In der Statistik-Literatur werden nun verschiedene Verfahren vorgeschlagen die AOB zu schätzen. Diese seien im folgenden allgemein vorgestellt und auf die wichtigsten vorgestellten nicht-parametrischen Verfahren angewendet.

4.4.1 Normalskalierungsregeln

Ein einfaches Verfahren besteht darin, die wahre Dichte als normalverteilt mit Varianz s^2 anzunehmen ([Scott 92], S. 55f.; [Wand & Jones 95], S. 60f.). Die Standardabweichung s wird dazu durch die Standardabweichung $\hat{s}$ der Stichprobe, oder falls möglich der Daten der Instanz

selbst, geschätzt. Diese Regel wird als *Normalskalierungsregel* (engl. *normal scale rule* oder *normal reference rule*) bezeichnet, die geschätzte Bandbreite als h_{ns} bzw. H_{ns}.

Eine Variante der Normalskalierungsregel besteht darin, anstatt der Standardabweichung entsprechend gewichtet den Interquartilsabstand zu nehmen. Der *Interquartilsabstand IR* ist definiert als die Differenz von 3/4-Quartil $x_{0,75}$ und 1/4-Quartil $x_{0,25}$. Dabei wird sich zunutze gemacht, daß für die Normalverteilung die folgende Beziehung gilt:

$$s = IR / (\Phi^{-1}(3/4) - \Phi^{-1}(1/4)) \approx IR / 1{,}348\,, \tag{4.34}$$

wobei mit $\Phi^{-1}(x)$ das x-Quantil der Standardnormalverteilung bezeichnet wird. Die so geschätzte Bandbreite sei mit h_{IR} bzw. H_{IR} bezeichnet. In praktischen Tests hat sich gezeigt, daß die Normalskalierungsregel mit der Standardabweichung häufig bei multimodalen Verteilungen überglättet und es sich bei der Normalskalierungsregel mit Interquartilsabstand um ein etwas robusteres Schätzverfahren für die Bandbreite handelt. Im Zweifelsfall empfiehlt es sich, die kleinere der beiden Bandbreiten zu wählen, vgl. [Scott 92].

Die Normalskalierungsregel mit beiden Varianten wird in den Experimenten in Kapitel 5 sowohl für den Equi-Width-Histogrammschätzer als auch für den Kernschätzer verwendet.

Normalskalierungsregeln bei Histogrammen

Für den univariaten Equi-Width-Histogrammschätzer ergibt sich mit $R(\Phi'(x, \mu, s)) = 1/(4\sqrt{\pi}s^3)$ aus Gleichung (4.27) die folgende Normalskalierungsregel

$$h_{EW\text{-}ns} = (24\sqrt{\pi})^{1/3} \cdot s \cdot n^{-1/3} \approx 3{,}4908 \cdot s \cdot n^{-1/3} \tag{4.35}$$

mit s der Standardabweichung der wahren Dichte f und n der Stichprobengröße.

In der Variante mit dem Interquartilsabstand lautet dies

$$h_{EW\text{-IR}} = \frac{(24\sqrt{\pi})^{1/3}}{(\Phi^{-1}(3/4) - \Phi^{-1}(1/4))} \cdot IR \cdot n^{-1/3} \approx 2{,}5896 \cdot IR \cdot n^{-1/3} \tag{4.36}$$

Sei nun im multivariaten Fall angenommen, daß $f(\boldsymbol{x}) = \Phi(\boldsymbol{x}, m, S)$ mit $S = Diag(s_1^2, \ldots, s_d^2)$.

Dann ergibt sich mit $R(f_j) = (2^{d+1}\pi^{d/2}s_j^2 s_1 \ldots s_d)^{-1}$ aus Gleichung (4.29) die folgende Normalskalierungsregel für den multivariaten Equi-Width-Histogrammschätzer.

$$h_{EW\text{-}ns,\,j} = 2^{\frac{d+1}{2}} \pi^{\frac{d}{4}} s_j (s_1 \ldots s_d)^{\frac{1}{2}} 6^{\frac{1}{d+2}} \prod_{i=1}^{d} \left(2^{d+1} \pi^{\frac{d}{2}} s_j^2 s_1 \ldots s_d \right)^{\frac{-1}{2(d+2)}} n^{\frac{-1}{d+2}}$$

$$= 2^{\frac{d+1}{2}} \pi^{\frac{d}{4}} s_j (s_1 \ldots s_d)^{\frac{1}{2}} 2^{\frac{1}{d+2}} 3^{\frac{1}{d+2}} \left(2^{d+1} \pi^{\frac{d}{2}} s_1 \ldots s_d \right)^{\frac{-d}{2(d+2)}} \prod_{i=1}^{d} s_j^{\frac{-1}{(d+2)}} n^{\frac{-1}{d+2}}$$

$$= 2 \cdot 3^{\frac{1}{d+2}} \pi^{\frac{d}{2(d+2)}} s_j (s_1 \ldots s_d)^{\frac{1}{d+2}} (s_1 \ldots s_d)^{\frac{-1}{d+2}} n^{\frac{-1}{d+2}}$$

$$= 2 \cdot 3^{\frac{1}{d+2}} \pi^{\frac{d}{2(d+2)}} s_j n^{\frac{-1}{d+2}} \tag{4.37}$$

Im bivariaten Fall ergibt sich daraus mit $d = 2$:

$$h_{EW\text{-}ns,\,j} = 2 \cdot 3^{1/4} \pi^{1/4} s_j n^{-1/4} \approx 3{,}504\; s_j n^{-1/4} \tag{4.38}$$

Normalskalierungsregeln beim Häufigkeitspolygonschätzer

Im univariaten Fall ergibt sich mit $R(\Phi''(x;\mu, s^2)) = 3/(8\sqrt{\pi} s^5)$ aus Gleichung (4.30) die folgende Normalskalierungsregel für den Häufigkeitspolygonschätzer:

$$h_{FP\text{-}ns} = 2\left(\frac{40\sqrt{\pi}}{49} \right)^{1/5} s n^{-1/5} \approx 2{,}1533\; s n^{-1/5} \tag{4.39}$$

Für multivariate Häufigkeitspolygonschätzer ist in [Scott 92], S. 109, folgende Regel angegeben:

$$h_j = 2 s_j n^{-1/(d+4)} \tag{4.40}$$

Normalskalierungsregeln beim Average Shifted Histogrammschätzer

Im univariaten Fall ergibt sich für den Average Shifted Histogrammschätzer mit $m \to \infty$ aus Gleichung (4.31) mit $R(\Phi''(x;\mu, s^2)) = 3/(8\sqrt{\pi} s^5)$:

$$h_{ASH,\,ns} = 2(2\sqrt{\pi})^{1/5} s n^{-1/5} \approx 2{,}5760\; s n^{-1/5} \tag{4.41}$$

Normalskalierungsregeln bei Kernschätzern

Wird der Epanechnikow-Kern benutzt, so ergibt sich aus Gleichung (4.33) mit $R(K_{Epa}) = 3/5$ und $R(\Phi''(x;\mu, s^2)) = 3/(8\sqrt{\pi}s^5)$ als Normalskalierungsregel für den univariaten Kernschätzer

$$h_{K\text{-}ns} = (40\sqrt{\pi})^{1/5} \cdot s \cdot n^{-1/5} \approx 2{,}3449 \cdot s \cdot n^{-1/5} \tag{4.42}$$

mit s der Standardabweichung von f und n der Stichprobengröße.

In [Scott 92], S. 152, ist folgende multivariate Normalskalierungsregel unter Verwendung des Gauß-Produktkerns angegeben:

$$h_{K\text{-}ns,\,j} = \left(\frac{4}{d+2}\right)^{1/(d+4)} \cdot s_j n^{-1/(d+4)} \tag{4.43}$$

Für andere Produktkerne K erhält man die entsprechende Bandbreite, indem man durch die Standardabweichung $\kappa_2^{1/2}(K)$ dieses Kernes dividiert.

Im bivariaten Fall ergibt sich mit $d = 2$ für den Gauß-Kern:

$$h_j = s_j n^{-1/6} \tag{4.44}$$

4.4.2 Cross-Validierung

In [Wand & Jones 95] z.B. wird die *Cross-Validierung* als ein weiteres bekanntes Verfahren zur Bestimmung der Bandbreite beschrieben. Idee ist dabei die Wiederverwendung von Daten. Es bezieht sich allerdings nicht wie die anderen in diesem Kapitel beschriebenen Verfahren auf den AMISE, sondern hat das Ziel den MISE selbst zu minimieren.

Betrachtet man den MISE des Schätzers $\hat{f}$, so erhält man:

$$\begin{aligned} MISE(f, \hat{f}, h) &= E\left(\int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 dx\right) \\ &= E\left(\int_{-\infty}^{\infty} \hat{f}(x)^2 dx\right) - 2E\left(\int_{-\infty}^{\infty} \hat{f}(x) f(x) dx\right) + \int_{-\infty}^{\infty} f(x)^2 dx \end{aligned} \tag{4.45}$$

Da der letzte Term nicht vom Schätzer $\hat{f}$ abhängt, genügt es den folgenden Ausdruck zu minimieren:

$$L(f,\hat{f},h) = E\left(\int_{-\infty}^{\infty}\hat{f}(x)^2dx\right) - 2E\left(\int_{-\infty}^{\infty}\hat{f}(x)f(x)dx\right) \tag{4.46}$$

Hier setzt die im folgenden vorgestellte Methode der Kleinste Quadrate Cross-Validierung (*least squares cross validation*) an ([Wand & Jones 95], S. 63f.). Für die Funktion $L(f,\hat{f},h)$ wird ein Schätzer gesucht und in Abhängigkeit von der Bandbreite *h* minimiert um einen optimalen Glättungsparameter zu erhalten. Da der zweite Term in der vorherigen Gleichung von der unbekannten wahren Dichte *f* abhängt, wird er durch einen geeigneten Schätzer ersetzt. Somit ergibt sich bei gegebener Stichprobe $X_1,...,X_n$ als eine erwartungstreue Schätzfunktion für die Funktion $L(f,\hat{f},h)$ die folgende Funktion in Abhängigkeit von der Bandbreite *h* (vgl. [Wand & Jones 95], S. 63f.):

$$\hat{L}(h) = \int_{-\infty}^{\infty}\hat{f}(x)^2dx - 2\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i) \text{ mit} \tag{4.47}$$

$$\hat{f}_{-i}(x) = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}K_h(x-X_j) \tag{4.48}$$

Das Verfahren besitzt jedoch gewisse Nachteile wie in [Wand & Jones 95] beschrieben wird. Die zu minimierende Funktion $L(f,\hat{f},h)$ kann mehrere lokale Minima besitzen, wobei das globale Minimum nicht immer dasjenige Minimum ist, welches das optimale *h* liefert. Für die einzelnen lokalen Minima muß entschieden werden, welche das optimale *h* liefert. Somit ist eine rein automatische Bestimmung der optimalen Bandbreite nicht möglich. Des weiteren sind zur Bestimmung der Bandbreite *h* nach dieser Methode sehr viel komplexere Berechnungen erforderlich, die dieses Verfahren nicht zur Verwendung auf Computern empfehlen. In [Wand & Jones 95], S. 86, werden für allgemeine Bandbreitenbestimmungen andere Verfahren wie die im nächsten Abschnitt vorgestellte Direkte Plug-In Methode empfohlen.

Aus diesen Gründen wird dieses Verfahren hier nicht weiter berücksichtigt. Empfehlenswerter sind die folgenden Verfahren [Wand & Jones 95].

4.4.3 Direkte Plug-In Verfahren

In [Wand & Jones 95], S. 71 f., wird das sogenannte *Direkte Plug-In Verfahren* für praktische Anwendungen empfohlen. Idee dieser Methode ist, Schätzungen der unbekannten wahren Dichte oder einer ihrer Ableitungen in den entsprechenden asymptotischen Ausdruck einzuset-

zen. Dies kann zu einem iterativen Verfahren ausgebaut werden, indem in einem Iterationsschritt die Schätzung der Dichtefunktion aus dem vorigen Iterationsschritt verwendet wird um die AOB zu schätzen. Mittels dieser Schätzung der AOB kann eine neue Schätzung der Dichtefunktion erfolgen. Im ersten Iterationsschritt kann z.B. die Normalskalierungsregel verwendet werden um eine erste Schätzung der Dichte zu erhalten. Der Einfluß der Normalskalierungsregel verschwindet so bei steigender Anzahl von Iterationsschritten. Nach *l* Iterationsschritten wird das Verfahren abgebrochen. Für die zu wählende Anzahl der Schritte sind allerdings aus der mathematischen Statistik keine näheren Erkenntnisse bekannt, [Wand & Jones 95], S. 71f., schlagen ein 2-Schritt-Verfahren vor. Die so gefundene Bandbreite wird im folgenden mit $h_{DPI\text{-}2}$ bezeichnet.

Um sich dieses Prinzip nutzbar zu machen, wird folgender funktionaler Zusammenhang ausgenutzt, der durch partielle Integration bei entsprechenden Voraussetzungen an *f* folgt, vgl. [Wand & Jones 95], S. 67f.:

$$R(f^{(s)}) = \int (f^{(s)}(x))^2 dx = (-1)^s \int f^{(2s)}(x) f(x) dx \tag{4.49}$$

Definiere nun das folgende Funktional

$$\psi_r = \int f^{(r)}(x) f(x) dx = E(f^{(r)}(X)) \text{ für } r \text{ gerade.} \tag{4.50}$$

Das Funktional läßt sich nun wiederum schätzen durch Verwendung eines Schätzers für die entsprechende Ableitung von *f*:

$$\hat{\psi}_r(h_r) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}^{(r)}(X_i, h_r) \tag{4.51}$$

Unter Verwendung eines Kernschätzers mit Kernfunktion *K* ergibt dies den Schätzer

$$\hat{\psi}_r(h_r) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{h_r}^{(r)}(X_i - X_j) \tag{4.52}$$

Für die weiteren Betrachtungen seien die folgenden Bedingungen vorausgesetzt, vgl. [Wand & Jones 95]:

- Die Kernfunktion *K* sei eine symmetrische Kernfunktion der Ordnung *k*, mit *k* = 2,4,..., und besitze *r* Ableitungen, wobei gelte $(-1)^{(r+k)/2+1} \cdot K^{(r)}(0) \cdot k_2(K) > 0$;
- Die wahre Dichte *f* sei *p*-mal stetig differenzierbar mit $p > k$;
- Für die Bandbreite $h_r = h = h(n)$ gelte $\lim_{n \to \infty} h = 0$ und $\lim_{n \to \infty} nh^{2r+1} = \infty$.

Dieser Schätzer kann nun in den entsprechenden Ausdruck für die AOB eingesetzt werden. Beim Kernschätzer z.B. ergibt sich somit aus Gleichung (4.33) die folgende Gleichung

$$\hat{h}_{K\text{-}DPI} = \left(\frac{R(K)}{k_2^2(K)\hat{\psi}_4(h_4)} \right)^{1/5} \cdot n^{-1/5} \tag{4.53}$$

Unglücklicherweise hängt die Gleichung (4.53) nun wiederum von der unbekannten Bandbreite h_4 ab. Diese kann nun bestimmt werden mittels einer asymptotischen Betrachtung des MSE des Schätzers ψ_r - zur Herleitung s. [Wand & Jones 95], S.67 f.:

$$h_r = \left(\frac{k!K^{(r)}(0)}{-k_2(K)\psi_{r+k}} \right)^{1/(r+k+1)} \cdot n^{-1/(r+k+1)}, \tag{4.54}$$

wobei K eine Kernfunktion der Ordnung k ist. Für h_4 aus Gleichung (4.53) ergibt sich unter Verwendung einer Kernfunktion der Ordnung 2:

$$h_4 = \left(\frac{2K^{(6)}(0)}{-k_2(K)\psi_6} \right)^{1/7} \cdot n^{-1/7} \tag{4.55}$$

Dieses Verfahren kann nun iterativ ausgebaut werden, denn zur Schätzung von ψ_6 ist zunächst der Bandbreitenparameter h_6 zu bestimmen und so weiter. [Wand & Jones 95], S. 71, schlagen ein 2-Schritt-Verfahren für praktische Anwendungen vor. Um ein anfängliches h zu bestimmen, läßt sich z.B. für Gleichung (4.51) analog zur Normalskalierungsregel die Annahme anwenden, daß es sich um eine normalverteilte Dichte mit Standardabweichung s handelt. Somit ergibt sich für ψ_r:

$$\psi_r = \frac{(-1)^{r/2} r!}{2^{r+1}(r/2)!\sqrt{\pi}} s^{-(r+1)} \tag{4.56}$$

Unter Verwendung der obigen Gleichungen läßt sich nun ein Algorithmus für das univariate 2-Schritt Direkte Plug-In Verfahren herleiten. Dabei werden abweichend von [Wand & Jones 95] die Gleichungen zur praktischen Berechnung anders eingesetzt ohne daß sich das Ergebnis ändert. Außerdem ist die resultierende Gleichung (4.60) für $h_{\text{DPI-2}}$ mit der Gleichung (4.42) für h_{NS} aus der Normalskalierungsregel besser vergleichbar.

Algorithmus 4.1 (*2-Schritt Direktes Plug-In Verfahren*):

Gegeben: Schätzung $\hat{s}$ der Standardabweichung der Dichte f und Stichprobe $X_1,...,X_n$, 6-fach differenzierbare Kernfunktion K der Ordnung 2.

1. Bestimme $\varphi_8 = \psi_8 s^9$ mittels Gleichung (4.56):

$$\varphi_8 = \frac{105}{32\sqrt{\pi}} \approx 1,851 . \tag{4.57}$$

2. Bestimme $\varphi_6(\hat{h}_6) = \psi_6(\hat{h}_6)s^7$ mittels Gleichung (4.52) und mit

$$\hat{h}_6 = \left(\frac{-2K_6(0)}{\mu_2(K)\varphi_8}\right)^{1/9} \cdot \hat{s} \cdot n^{-1/9} \tag{4.58}$$

3. Bestimme $\varphi_4(\hat{h}_4) = \psi_4(\hat{h}_4)s^5$ mittels Gleichung (4.52) und mit

$$\hat{h}_4 = \left(\frac{-2K_4(0)}{\mu_2(K)\varphi_6(\hat{h}_6)}\right)^{1/7} \cdot \hat{s} \cdot n^{-1/7} \tag{4.59}$$

4. Die Bandbreite $h_{DPI\text{-}2}$ wird nun geschätzt mittels

$$\hat{h}_{DPI\text{-}2} = \left(\frac{R(K)}{\mu_2(K)^2\varphi_4(\hat{h}_4)}\right)^{1/5} \cdot \hat{s} \cdot n^{-1/5} \tag{4.60}$$

Man beachte daß für das Direkte Plug-In Verfahren Kerne höherer Ordnung mit Existenz höherer Ableitung erforderlich ist. Für das 2-Schritt Direktes Plug-In Verfahren ist ein Kern mindestens der Ordnung 2 und 6-fach differenzierbar erforderlich.

Für die Experimente in Kapitel 5.5 wurde das 2-Schritt Direkte Plug-In Verfahren zur Schätzung der AOB eines univariaten Kernschätzers implementiert. Als Kernfunktion L_g wurde der Gaußkern gewählt, einer Kernfunktion 2-ter Ordnung, für die beliebig viele Ableitungen existieren.

4.4.4 Verfahren der variablen Bandbreite

Grundidee der variablen oder adaptiven Bandbreite ist, daß in Bereichen hoher Dichte eine kleinere Bandbreite erforderlich ist als in Bereichen mit niedriger Dichte. Um die Bandbreite der lokalen Dichte anzupassen, existieren grundsätzlich zwei verschiedene Ansätze. Im dem einen Ansatz wird die Bandbreite in Abhängigkeit von den Stichprobenwerten X_i gewählt: $h = h_i = h(X_i)$. In dem anderen Ansatz hängt die Bandbreite von der Position des Punktes x ab, an dem die Dichte geschätzt werden soll: $h = h(x)$. Die Wahl des Ansatzes hängt von praktischen Anforderungen ab, vgl. dazu z.B. [Scott 92], S. 181f.

Im folgenden wird ein Ansatz der adaptiven Bandbreitenbestimmung in Abhängigkeit von der Stichprobe nach [Silverman 86], S. 101, vorgestellt.

Algorithmus 4.2 (*adaptive Bandbreitenbestimmung*)

Gegeben: Stichprobe $X_1,...,X_n$

1. Finde eine erste Schätzung $f_0(t)$, die die Bedingung $f_0(X_i)>0$ für alle i erfüllt.

2. Definiere lokale Bandbreitenfaktoren λ_i mittels

$$\lambda_i = \left(\frac{g}{\tilde{f}(X_i)} \right)^{\alpha} \tag{4.61}$$

wobei g das geometrische Mittel der $\tilde{f}(X_i)$ ist:

$$\log g = \frac{1}{n} \sum_{i=1}^{n} \log \tilde{f}(X_i) . \tag{4.62}$$

α ist ein Empfindlichkeitsparameter mit $0 \le \alpha \le 1$.

3. Die adaptive Bandbreite ist nun definiert durch $h_i = h(X_i) = h\lambda_i$.

In den experimentellen Untersuchungen dieser Arbeit wurde die adaptive Bandbreitenbestimmung für den univariaten Kernselektivitätsschätzer implementiert. Für die erste Schätzung f_0 wurde ein Kernschätzer mit fester Bandbreite verwendet, die nach der Normalskalierungsregel bestimmt wurde. Der Parameter α wurde in Übereinstimmung mit [Silverman 86] als $\alpha = 1/2$ gewählt. Die Ergebnisse sind in Kapitel 5.5 vorgestellt und diskutiert.

Ein weiteres Verfahren, das ebenfalls in den Bereich der variablen Bandbreitenbestimmung fällt, besteht darin, die Stichprobe mittels einer zu einer bestimmten Transformationsfamilie gehörenden Funktion abzubilden, anschließend die Dichte der transformierten Stichprobe zu schätzen und zuletzt die Dichte zurück zu transformieren. Diese Methode kann insbesondere dann angewendet werden, wenn bekannt ist, daß die Dichte der zugrundeliegenden Daten zu einer bestimmten Verteilungsfamilie gehört. So können z.B. die exponential verteilten Daten durch geeignete logarithmische Funktionen transformiert werden. Für weitere Details zu diesem Ansatz siehe z.B. [Silverman 86] und [Wand et al. 91].

Aber auch der in Kapitel 3.5 vorgestellte Hybridschätzer stellt ein Verfahren der variablen Bandbreite dar, da in jedem Bin i.a. eine andere Bandbreite für den auf das Bin beschränkten Kernschätzer gewählt wird. Hauptvorteil des Hybridschätzer liegt jedoch im wesentlichen in der Berücksichtigung von Sprungstellen.

4.4.5 Besonderheiten der Verteilung

Die Schätzung der Bandbreite kann weiterhin verfeinert werden, in dem weitere Kennwerte in die Bestimmung der Bandbreite mitaufgenommen werden. [Scott 92] hat den Einfluß des Korrelationskoeffizienten, der Kurtosis und der Schiefe auf die Bandbreitenbestimmung mit der Normalskalierungsregel untersucht.

Korrelierte Daten

In den bisherigen Betrachtungen zur Bestimmung der AOB wurde von voneinander unabhängigen und damit unkorrelierten Daten ausgegangen. Im Falle von Korrelationen bekommt man eine bessere Schätzung des AOB unter Berücksichtigung des Korrelationsfaktors ρ.

In [Scott 92] wurde der bivariate Fall betrachtet. Geht man davon aus, daß es sich bei der wahren Dichte *f* um eine normalverteilte Dichte mit Randvarianzen s_1, s_2 sowie Korrelation ρ handelt, so folgt aus Gleichung (4.29) mit $R(f_1) = 1 \Big/ \left(8\pi(1-\rho^2)^{3/2} s_1^3 s_2 \right)$ ($R(f_2)$ analog):

$$h_i = 2 \cdot (3\pi)^{1/4} s_i (1-\rho^2)^{3/8} n^{-1/4}, \; i = 1, 2 \tag{4.63}$$

Interessant ist hierbei insbesondere der Ausdruck $(1-\rho^2)$. Sind die Daten unkorreliert ($\rho = 0$), so wird der Term gleich Eins und es ergibt sich für h_i die bekannte Gleichung (4.38). Je stärker die Daten jedoch korreliert sind, desto mehr nähert sich der Term dem Wert Null. Dies führt dazu, daß auch die AOB gegen Null geht, während der AMISE (s.u.) gegen unendlich strebt. Der AMISE unter den obigen Voraussetzungen ergibt sich als (vgl. unten Gleichung (4.68)):

$$\text{AMISE} = \frac{3}{2}(48\pi)^{-\frac{1}{2}} \frac{1}{s_1 s_2}(1-\rho^2)^{-3/4} n^{\frac{-1}{2}} \approx 0,122 \frac{1}{s_1 s_2}(1-\rho^2)^{-3/4} n^{\frac{-1}{2}} \tag{4.64}$$

Die Bestimmung der Bandbreite beim Kernschätzer unter Berücksichtigung des Korrelationsfaktors erfolgt analog, vgl. [Scott 92].

Daten mit hoher Kurtosis oder Schiefe

Des weiteren ist es möglich, Verteilungen mit speziellen Eigenschaften wie Kurtosis oder Schiefe gesondert zu behandeln. Für solche Verteilungen existieren in der Literatur spezielle Faktoren zur Wahl der asymptotisch optimalen Bandbreite, vgl. dazu [Scott 92]. Diese Faktoren sind jedoch sehr kompliziert für den Einsatz in praktischen Anwendungen. Für den univariaten Histogrammschätzer seien im folgenden die in [Scott 92] berechneten Faktoren für schiefe Verteilungen bzw. solche mit hoher Kurtosis angegeben. Diese Faktoren werden bei der Normal-

skalierungsregel angewendet. Die Idee ist, anstelle der Normalverteilung entsprechend andere Verteilungen zugrunde zu legen.

Für eine schiefe Verteilung wird in [Scott 92], S. 56, der folgende Faktor $\varphi_{schiefe}$ - hergeleitet auf Grundlage der Lognormal-Verteilung - vorgeschlagen:

$$\varphi_{schiefe}(s) = \frac{2^{1/3} s}{e^{5s^2/4}(s^2+2)^{1/3}\left(e^{s^2}-1\right)^{1/2}} \tag{4.65}$$

Für die Kurtosis wird in [Scott 92], S. 57, der folgende Faktor $\varphi_{kurtosis}$- hergeleitet auf Grundlage der *t*-Verteilung mit $\nu>1$ Freiheitsgraden - vorgeschlagen:

$$\varphi_{kurtosis}(\nu) = \frac{(\nu-2)^{1/2} B\left(\frac{1}{2}, \frac{\nu+1}{2}\right)^{2/3}}{2^{2/3}\pi^{1/6}(\nu+1)^{2/3} B\left(\frac{3}{2}, \frac{2\nu+3}{2}\right)^{1/3}}, \; B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{4.66}$$

Für die Definition der Gamma-Funktion Γ vergleiche z.B. [Forster 79].

4.5 Konvergenzordnung bei (bzgl. des AMISE) optimaler Bandbreite

Mit Hilfe der im vorigen Abschnitt bestimmten asymptotisch optimalen Bandbreiten lassen sich Aussagen über die Konvergenzordnung der Schätzer treffen, indem die Bandbreiten in die jeweiligen Gleichungen für den AMISE eingesetzt werden.

Konvergenzordnung beim Histogrammschätzer

Im univariaten Fall ergibt sich durch Einsetzen von Gleichung (4.27) in Gleichung (4.7) für den Equi-Width-Histogrammschätzer:

$$AMISE(h_{AO}) = n^{-1}h_{AO}^{-1} + \frac{1}{12}h_{AO}^{\;2}R(f^{(1)})$$

$$= n^{-1}\left(\frac{6}{R(f^{(1)})}\right)^{-1/3} n^{1/3} + \frac{1}{12}\left(\left(\frac{6}{R(f^{(1)})}\right)^{1/3} \cdot n^{-1/3}\right)^2 R(f^{(1)})$$

$$= \left(\frac{6}{R(f^{(1)})}\right)^{-1/3} n^{-2/3} + \frac{1}{2}\left(\frac{6}{R(f^{(1)})}\right)^{2/3} \cdot n^{-2/3}\frac{R(f^{(1)})}{6}$$

$$= \left(\frac{6}{R(f^{(1)})}\right)^{-1/3} n^{-2/3} + \frac{1}{2}\left(\frac{6}{R(f^{(1)})}\right)^{-1/3} \cdot n^{-2/3}$$

$$= \frac{3}{2}\left(\frac{6}{R(f^{(1)})}\right)^{-1/3} n^{-2/3} = O(n^{-2/3}). \tag{4.67}$$

Im multivariaten Fall ergibt sich durch Einsetzen von Gleichung (4.29) in Gleichung (4.9) für den Equi-Width-Histogrammschätzer:

$$AMISE(h_{AO}) = \left(n \prod_{i=1}^{d} h_{AO,i}\right)^{-1} + \frac{1}{12} \sum_{i=1}^{d} h_{AO,i}^2 R(f_i)$$

$$= \left(n \prod_{j=1}^{d} \left(R(f_j)^{-\frac{1}{2}} 6^{\frac{1}{d+2}} \left(\prod_{i=1}^{d} R(f_i)^{\frac{1}{2(d+2)}}\right) n^{\frac{-1}{d+2}}\right)\right)^{-1}$$
$$+ \frac{1}{12} \sum_{j=1}^{d} \left(R(f_j)^{-\frac{1}{2}} 6^{\frac{1}{d+2}} \left(\prod_{i=1}^{d} R(f_i)^{\frac{1}{2(d+2)}}\right) n^{\frac{-1}{d+2}}\right)^2 R(f_j)$$

$$= n^{\frac{-2}{d+2}} 6^{\frac{-d}{d+2}} \prod_{j=1}^{d} \left(R(f_j)^{\frac{1}{2}} \prod_{i=1}^{d} R(f_i)^{\frac{-1}{2(d+2)}}\right) + \frac{1}{2} 6^{\frac{-d}{d+2}} \left(\prod_{i=1}^{d} R(f_i)^{\frac{1}{d+2}}\right) n^{\frac{-2}{d+2}}$$

$$= 6^{\frac{-d}{d+2}} \left(\prod_{j=1}^{d} \left(R(f_j)^{\frac{1}{2}} \prod_{i=1}^{d} R(f_i)^{\frac{-1}{2(d+2)}}\right) + \frac{1}{2}\left(\prod_{i=1}^{d} R(f_i)^{\frac{1}{d+2}}\right)\right) n^{\frac{-2}{d+2}}$$

$$= \frac{3}{2} 6^{\frac{-d}{d+2}} \left(\prod_{i=1}^{d} (R(f_i))^{\frac{1}{d+2}}\right) n^{\frac{-2}{d+2}} \tag{4.68}$$

$$= O(n^{-2/(d+2)}). \tag{4.69}$$

Konvergenzordnung beim Häufigkeitspolygon

Im univariaten Fall ergibt sich durch Einsetzen von Gleichung (4.30) in Gleichung (4.10) für den Häufigkeitspolygonschätzer:

$$\mathrm{AMISE}(h_{AO}) = \frac{2}{3} n^{-1} h^{-1} + \frac{49}{2880} h^4 R(f^{(2)})$$

$$= \frac{2}{3}n^{-1}\frac{1}{2}\left(\frac{15}{49R(f^{(2)})}\right)^{-1/5}n^{1/5} + \frac{49}{2880}\left(2\left(\frac{15}{49R(f^{(2)})}\right)^{1/5}n^{-1/5}\right)^4 R(f^{(2)})$$

$$= \left(\frac{1}{3}\left(\frac{15}{49R(f^{(2)})}\right)^{-1/5} + \frac{49^{\frac{1}{5}}}{12}\left(\frac{R(f^{(2)})}{15}\right)^{\frac{1}{5}}\right)n^{-4/5}$$

$$= \frac{5}{12}\left(\frac{15}{49R(f^{(2)})}\right)^{-1/5}n^{-4/5}$$

$$= O(n^{-4/5}) \tag{4.70}$$

Im multivariaten Fall ist die Ordnung des Häufigkeitspolygonschätzers in [Scott 92], S. 107, angegeben als

$$\text{AMISE}(h_{AO}) = O(n^{-4/(d+4)}) \tag{4.71}$$

Konvergenzordnung beim Average Shifted Histogramm

Im univariaten Fall ergibt sich durch Einsetzen von Gleichung (4.31) in Gleichung (4.12) für den Average Shifted Histogrammschätzer für $m \to \infty$:

$$AMISE(h_{AO}) = \frac{2}{3}n^{-1}h^{-1} + \frac{1}{144}h^4 R(f^{(2)})$$

$$= \frac{2}{3}n^{-1}\left(\frac{24}{R(f^{(2)})}\right)^{-1/5}n^{1/5} + \frac{1}{144}\left(\left(\frac{24}{R(f^{(2)})}\right)^{1/5}n^{-1/5}\right)^4 R(f^{(2)})$$

$$= \left(\frac{2}{3}\left(\frac{24}{R(f^{(2)})}\right)^{-1/5} + \frac{1}{6}\left(\frac{24}{R(f^{(2)})}\right)^{-1/5}\right)n^{-4/5}$$

$$= \frac{5}{6}\left(\frac{1}{24}R(f^{(2)})\right)^{1/5}n^{-4/5}$$

$$= O(n^{-4/5}) \tag{4.72}$$

Im multivariaten Fall ist die Ordnung des Average Shifted Histogrammschätzers in [Scott 92], S. 121, angegeben als

$$\text{AMISE}(h_{AO}) = O(n^{-4/(d+4)}) \tag{4.73}$$

Konvergenzordnung beim Kernschätzer

Im univariaten Fall ergibt sich durch Einsetzen von Gleichung (4.33) in Gleichung (4.20) für den Kernschätzer:

$$\begin{aligned}
AMISE(h_{AO}) &= n^{-1}h_{AO}^{-1}R(K) + h_{AO}^4\frac{k_2^2}{4}(K)R(f^{(2)}) \\
&= n^{-1}\left(\frac{R(K)}{k_2^2(K)R(f^{(2)})}\right)^{\frac{-1}{5}} n^{\frac{1}{5}}R(K) + \left(\frac{R(K)}{k_2^2(K)R(f^{(2)})}\right)^{\frac{4}{5}} n^{\frac{-4}{5}}\frac{k_2^2(K)}{4}R(f^{(2)}) \\
&= \left((k_2^2(K)R(f^{(2)}))^{\frac{1}{5}}(R(K))^{\frac{4}{5}} + \frac{1}{4}(R(K))^{\frac{4}{5}}(k_2^2(K)R(f^{(2)}))^{\frac{1}{5}}\right)n^{-\frac{4}{5}} \\
&= \frac{5}{4}(R(K))^{\frac{4}{5}}(k_2^2(K)R(f^{(2)}))^{\frac{1}{5}}n^{-\frac{4}{5}} \\
&= O(n^{-4/5}).
\end{aligned} \tag{4.74}$$

Im multivariaten Fall ergibt sich für die Konvergenzordnung des Produktkernschätzers:

$$\text{AMISE}(h_{AO}) = O(n^{-4/(d+4)}) \tag{4.75}$$

Zusammengefaßt seien die Konvergenzordnungen der verschiedenen Schätzer bzgl. des AMISE noch einmal in der folgenden Übersicht dargestellt.

Schätzer	**univariat**	**multivariat**
empirische Verteilungsfunktion	$O(n^{-1/2})$	$O(n^{-1/(d+1)})$
Equi-Width-Histogrammschätzer	$O(n^{-2/3})$	$O(n^{-2/(d+2)})$
Häufigkeitspolygonschätzer	$O(n^{-4/5})$	$O(n^{-4/(d+4)})$

Tabelle 4.5: Konvergenzordnung der verschiedenen Schätzer bzgl. des AMISE

Schätzer	univariat	multivariat
Average Shifted Histogrammschätzer	$O(n^{-4/5})$	$O(n^{-4/(d+4)})$
Kernschätzer	$O(n^{-4/5})$	$O(n^{-4/(d+4)})$

Tabelle 4.5: Konvergenzordnung der verschiedenen Schätzer bzgl. des AMISE

Unter der Voraussetzung einer asymptotisch optimalen Bandbreite sind die zwei wesentlichen Parameter, die den AMISE und damit die Konvergenzordnung beeinflussen, die Dimension d und die Stichprobengröße n. Ihr Einfluß wird in den nächsten Abschnitten besprochen.

4.6 Das Dimensionenproblem

Als ein die Konvergenzordnung wesentlich beeinflussender Parameter wurde die Dimension d identifiziert. Abbildung 4.2 zeigt die variablen Terme der Konvergenzordnung (rote Kurve $n^{-2/(d+2)}$, blaue gepunktete Kurve $n^{-4/(d+4)}$) bei fester Stichprobengröße $n = 1000$ und wachsender Dimension für die verschiedenen Schätzverfahren (Die anderen Terme hängen weder von n noch von d ab). Wie zu erwarten (offensichtlich gilt $\lim_{d \to \infty} n^{-c/(d+c)} = 1$ sowie $n^{-c/((d+1)+c)} > n^{-c/(d+c)}$ für $c = 2, 4$ und $n > 0$ bel. aber fest), steigen die Terme monoton an und streben gegen den Wert 1. Gerade am Anfang ist der Anstieg der Kurve sehr steil, so daß hier sehr schnell höhere Fehler zu erwarten sind.

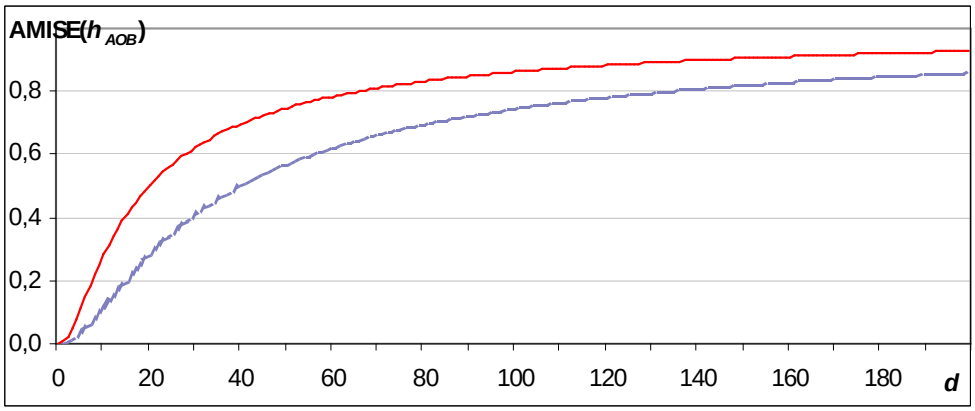


Abbildung 4.2: Zur Konvergenzordnung bei fester Stichprobengröße und variabler Dimension.

Diese mit steigender Dimension höheren Fehler sind nur durch eine größere Stichprobe zu kompensieren. Um bei Erhöhung der Dimension von d_1 auf $d_2 > d_1$ den gleichen AMISE zu erhalten, lassen sich die beiden Gleichungen mit unterschiedlicher Stichprobengröße n_1 und n_2 sowie

unterschiedlicher Dimension gleichsetzen und nach der zu bestimmenden Stichprobengröße n_2 auflösen. Dazu seien im folgenden c = 2, 4 und $n_1, n_2 > 1$ und $d_1, d_2 > 0$, alles ganze Zahlen:

$$AMISE(d_1, n_1) \cong AMISE(d_2, n_2) \Leftrightarrow$$

$$n_1^{-c/(d_1+c)} \cong n_2^{-c/(d_2+c)} \Leftrightarrow$$

$$n_2 \cong n_1^{(d_2+c)/(d_1+c)} \tag{4.76}$$

Tabelle 4.6 zeigt als Beispiel bei einer Ausgangsstichprobengröße von $n_1 = 100$ und Dimension $d_1 = 1$, wie schnell die erforderliche Stichprobengröße n_2 anwächst, um bei steigender Dimension d_2 den gleichen AMISE (hier mit $c = 4$) zu erhalten.

d_1	n_1	d_2	n_2
1	100	1	100
1	100	2	251,2
1	100	3	631,0
1	100	4	1.584,9
1	100	10	398.107,2
1	100	100	3,98107E+41

Tabelle 4.6: Zum Verhältnis Stichprobengröße und Dimension beim AMISE

4.7 Bestimmung der Stichprobengröße bei vorgegebener Fehlerschranke

Als ein weiterer die Konvergenzordnung wesentlich beeinflussender Parameter wurde die Stichprobengröße n identifiziert. Abbildung 4.3 zeigt die Konvergenzordnung bei Dimension $d = 2$ und wachsender Stichprobengröße n für die verschiedenen Schätzverfahren (rote Kurve $n^{-2/(d+2)}$, blaue gepunktete Kurve $n^{-4/(d+4)}$). Experimente mit uni- und bivariaten Testdaten bestätigen diese Ergebnisse, vgl. dazu Kapitel 5.4.1.

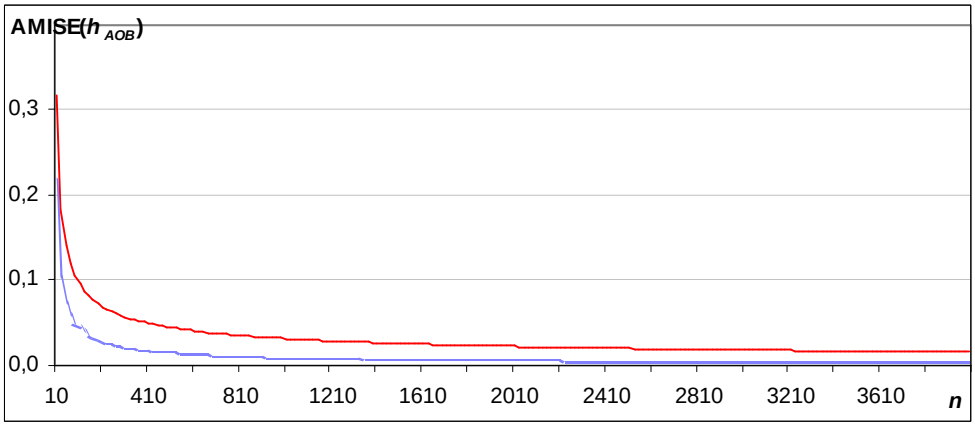


Abbildung 4.3: Zur Konvergenzordnung bei fester Dimension und variabler Stichprobengröße.

Interessant ist daher die Berechnung der erforderlichen minimalen Stichprobengröße unter Vorgabe eines gewünschten maximalen AMISE. Die Berechnung erfolgt analog zum vorigen Abschnitt, indem diesmal der AMISE einem gegebenen konstanten Wert A gleich gesetzt und bei gegebener Dimension d nach der Stichprobengröße n aufgelöst wird. Dazu seien wieder im folgenden $c = 2, 4$ und $n > 1$ und $d > 0$, alles ganze Zahlen:

$$AMISE(d, n) = A \Leftrightarrow$$

$$B \cdot n^{-c/(d+c)} = A \Leftrightarrow \quad \text{(mit } B \text{ = const. für das jeweilige Schätzverfahren)}$$

$$n = \left(\frac{A}{B}\right)^{-(d+c)/c} \tag{4.77}$$

Tabelle 4.7 zeigt die benötigte Stichprobengröße n bei fester Dimension $d = 2$ und dem erwarteten Wert von A/B mit $c = 4$. Dabei wurde der Ausgangswert von 0,1 jeweils halbiert und auf 4 Stellen nach dem Komma gerundet.

A/B	***n***
0,1000	32
0,0500	90
0,0250	253
0,0125	716
0,0063	2.024
0,0031	5.725
0,0016	16.191

Tabelle 4.7: Zur Abhängigkeit der Stichprobengröße n vom AMISE

A/B	***n***
0,0008	45.795
0,0004	129.527
0,0002	366.358
0,0001	1.036.216

Tabelle 4.7: Zur Abhängigkeit der Stichprobengröße *n* vom AMISE

In [Scott 92], S. 198f., wurde ebenfalls die Wahl einer äquivalenten Stichprobengröße bei variierender Dimension diskutiert. Diese Diskussion bezieht sich auf eine Betrachtung des MISE.

5. Experimente

In diesem Abschnitt werden ausführlich die Ergebnisse der Experimente mit verschiedenen implementierten Selektivitätsschätzern für Bereichsanfragen vorgestellt. Ziel der Experimente ist es, die theoretischen Betrachtungen der vorherigen Kapitel in der Praxis zu überprüfen. Dazu gehören die Dimension und der Einfluß der Stichprobengröße sowie die verschiedenen Verfahren zur Bestimmung eines optimalen Glättungsparameters. Des weiteren werden die verschiedenen Selektivitätsschätzer miteinander verglichen. Bei den Kernschätzern ist der Randbehandlung besondere Beachtung gewidmet.

5.1 Notation

In der folgenden Tabelle ist die in diesem Kapitel verwendete Notation angegeben:

Dimension	d
Domäne	D
bestimmt die Domäne: $D = [0, 2^p\text{-}1]$	p
linker Rand der Domäne (hier $l = 0$) rechter Rand (hier $r = 2^p\text{-}1$)	l r
Anzahl Testdatensätze (Größe der Grundgesamtheit)	N
(ganzzahlige) Testdaten der Größe N aus der Domäne$[0, 2^p\text{-}1]$ (univariat) bzw. aus $[0, 2^{p_1} - 1] \times [0, 2^{p_2} - 1]$ (bivariat)	$T = T(p,N)$ $T = T(p_1 \times p_2, N)$
Stichprobe zu $T(p)$ (bzw. $T(p_1,p_2)$) der Größe n	$S = S(T,p,..,n)$
untere Grenze einer einzelnen Fenster-Bereichsanfrage obere Grenze einer einzelnen Fenster-Bereichsanfrage einzelne (Fenster-) Bereichsanfrage	a b $Q = Q(a,b)$
prozentuale Anfragegröße, hier 1%, 2%, 5% oder 10% der Domänengröße	q
Anzahl Testanfragen	m
Anfragemenge zu $T(p)$ (bzw. $T(p_1,p_2)$) der Größe r mit Anfragen der Größe q	$A = A(T,p,..,q,m)$
Bandbreite bzw. Anzahl Histogramm-Bins	h, k
Verfahren zur Best. der Bandbreite zu T bzw. S, s.u.	$B = B(T,N,S,n,h_0)$
Selektivitätsschätzverfahren zu Anfragen $Q(T,p,q,r)$ auf Stichprobe $S = S(T,p,..,n)$ der Testdaten $T(p,N)$ (bzw. $T(p_1,p_2)$) mit Bandbreite h bzw. Anzahl Bins k mittels b generiert, s.u.	$y = y(B) = y(T,p,..,N,S,n,Q,q,m,h\|k)$
Anzahl Shifts bei Average Shifted Histogramm	s
beim Kernselektivitätsschätzer verwendete Kernfunktion; falls nichts anderes gesagt ist, ist $K = K_{Epa}$ der Epanechnikow-Kern bzw. für d > 1 ist $K = K_{Epa}^P$ der Epanechnikow-Produktkern.	K

Tabelle 5.1: Notation der in den Experimenten verwendeten Datenmengen, Verfahren und Parameter.

optimale (durch Experimente) gefundene Binweite bzw. Bandbreite (geschätzte) asymptotisch optimale Bandbreite	OGB AOB
absoluter bzw. relativer Selektivitätsfehler mittlerer ASF bzw. RSF	ASF, RSF MASF, MRSF

Tabelle 5.1: Notation der in den Experimenten verwendeten Datenmengen, Verfahren und Parameter.

Tabelle 5.2 gibt einen Überblick über die zur Schätzung der Selektivität in den folgenden Experimenten verwendeten Verfahren:

Verfahren	**Kürzel**	**Dim.**	**Parameter**
Gleichverteilungsannahme	GS	1,2	*d*, *D*, *A*(*q*), *m*
direkter Selektivitätsschätzer	DS	1,2	*d*, *S*(*p*,..), *n*, *A*(*q*), *m*
Histogrammselektvitätsschätzer Equi-Width-Histogrammselektivitätsschätzer Equi-Depth-Histogrammselektivitätsschätzer Max-Diff-Histogrammselektivitätsschätzer KD-Baum-Selektivitätsschätzer	HS EW-HS ED-HS MD-HS KDBS	1,2 1,2 1,2 1 2	*d*, *S*(*p*,..), *n*, *A*(*q*), *m*, *k*
Average Shifted Histogrammselektivitätsschätzer	ASHS	1,2	*d*, *S*(*p*,..), *n*, *A*(*q*), *m*, *k*, *s*
Histogrammselektivitätsschätzer auf Z-Ordnung - mit Equi-Width-Histogramm - mit Equi-Depth-Histogramm - mit Max-Diff-Histogramm	ZHS EW-ZHS ED-ZHS MD-ZHS	2 2 2 2	*d*, *S*(*p*,..), *n*, *A*(*q*), *m*, *k*
Kernselektivitätsschätzer Kernselektivitätsschätzer ohne Randbehandlung Kernselektivitätsschätzer mit Spiegelung Kernselektivitätsschätzer mit Rankkern	KS KS-O KS-S KS-R	1,2 1,2 1,2 1,2	*d*, *S*(*p*,..), *n*, *A*(*q*), *m*, *h*, *K*
Hybridselektivitätsschätzer	YS	1	*S*(*p*), *n*, *A*(*q*), *m*, *k*, *K*

Tabelle 5.2: Verwendete Selektivitätsschätzverfahren *y*

Die Ergebnisse der Gleichverteilungsannahme und des direkten Selektivitätsschätzers dienen als in den folgenden Experimenten als Referenzwerte für die Ergebnisse der anderen untersuchten Schätzer. Histogrammselektivitätsschätzer sind die in DBMS zur Zeit aktuell verwendeten (z.B. Oracle 8) und in der Literatur am meisten diskutierten Verfahren. Die für alle betrachteten Dimensionen klassischen Verfahren sind der Equi-Width- und der Equi-Depth-Histogrammselektivitätsschätzer, wobei im bivariaten Fall die Variante von [Muralikrishna & DeWitt 88] implementiert wurde. Im univariaten Fall wird weiterhin der von [Poosala et al. 96] vorgeschlagene Max-Diff-Histogrammselektivitätsschätzer untersucht. Diese univariaten Histogrammselektivitätsschätzer werden auch im multivariaten Fall auf die mittels einer Z-Ordnung abgebildeten Daten angewendet. An weiteren Verfahren werden der Average Shifted Histogrammselektivitätsschätzer, der Kernselektivitätsschätzer mit und ohne Randbehandlung sowie der Hybridselektivitätsschätzer untersucht.

Tabelle 5.3 gibt einen Überblick über die zur Schätzung der asymptotisch optimalen Bandbreite (AOB) verwendeten Verfahren:

Verfahren	Kürzel	Dim.	Parameter
Normalskalierunsregel	NS	1,2	*d, p, .., T, N, n*
direktes Plug-In Verfahren 2. Stufe	DPI-2	1	*p, T, N,* S, *n, K*
adaptives Verfahren	AD	1	*p, T, N,* S, *n, K*

Tabelle 5.3: Verwendete AOB-Schätzverfahren *B*

Bei der Normalskalierungsregel handelt es sich um ein einfaches Standardverfahren zur Schätzung der AOB. Sie wird daher sowohl für den Equi-Width-Histogrammselektivitätsschätzer als auch für den Kernselektivitätsschätzer verwendet. Dabei wird immer das Minimum der beiden Varianten mit Standardabweichung bzw. Interquartilsabstand gewählt, vgl. Kapitel 4.4.1. Das direkte Plug-In Verfahren 2. Stufe und das adaptive Verfahren werden als Erweiterungen ausführlich beim univariaten Kernselektivitätsschätzer untersucht.

5.2 Verwendete Fehlermaße

Zur Bewertung der Experimente werden der (mittlere) absolute Selektivitätsfehler bzgl. der Instanz - ASF (MASF) - sowie der (mittlere) relative Selektivitätsfehler - RSF (MRSF) - aus Kapitel 2.2 verwendet (vgl. Gleichungen (2.21) bzw. (2.24) für den ASF bzw. MASF sowie Gleichungen (2.22) bzw. (2.25) für den RSF bzw. MRSF). Im allgemeinen folgen die beiden Werte einander. Allerdings kann es in Bereichen mit sehr geringer Dichte zu einem höheren relativen Fehler kommen, da dort ein evtl. sehr kleiner absoluter Fehler durch eine sehr kleine wahre Selektivität dividiert wird. Beispielweise ergibt sich bei $N = 100.000$, $s = 2$ und $\ddot{\sigma} = 3$ ein absoluter Fehler bzgl. N von $\eta = 1/100.000 = 0{,}001\%$ und ein relativer Fehler von $\varepsilon = 1/3 = 33{,}333\%$, während sich bei $\sigma = 2000$ und $\hat{\sigma} = 2001$ auch ein absoluter Fehler bzgl. N von $\eta = 0{,}001\%$ ergibt aber ein relativer Fehler von $\varepsilon = 1/3000 = 0{,}033\%$. In den folgenden Abschnitten ist i.a. nur der (M)RSF wiedergegeben.

5.3 Testumgebung

Es folgt eine ausführliche Beschreibung der Testumgebung, d.h. insbesondere über die Auswahl und Erstellung der Testdaten, ihren statistischen Eigenschaften sowie über die Durchführung der Experimente.

5.3.1 Auswahl der Testdaten

Für die Experimente wurden sowohl künstliche als auch reale Datensätze benutzt. Als typische künstliche Datensätze wurden gleichverteilte, standard-normalverteilte und exponentialverteilte Daten ausgesucht und mittels Algorithmen aus [Knuth 69] erzeugt. Exponentialverteilte Daten zeichnen sich durch eine extreme Schiefe aus mit einer hohen Dichte am unteren

und einer geringen Dichte am oberen Rand, wie sie auch bei schiefen Zipf-verteilten Daten auftreten kann. Letztere findet sich häufig in der Literatur über Datenbanken als Vergleichsverteilung ([Mannino et al. 88], [Ioannidis & Poosala 95]). Als reale Datensätze dienten verschiedene Datensätze des U.S. Census Bureau, die im TIGER/Line-Format vorlagen, und via Internet veröffentlicht wurden. Hauptbestandteil der Originaldatensätze sind Linienstücke, wovon Anfangs- und Endpunkt als Datenbasis für die Experimente zur Selektivitätsschätzung extrahiert wurden. Es handelt sich daher um zwei-dimensionale Datensätze. Zur Evaluation der eindimensionalen Selektivitätsschätzer wurde jeweils nur eine Komponente der Datensätze betrachtet. Diese Datensätze zeichnen sich des weiteren durch eine sehr inhomogene Struktur aus und eignen sich daher sehr gut als Testdatensätze für die verwendeten Selektivitätsschätzer. Es wurden drei Datensätze mit möglichst großer Anzahl von Tupeln ausgesucht (28.136 Los-Angeles Daten, 52.120 Arapahoe-Daten und 257.942 Rail-Road-Daten).

Aus Implementierungsgründen wurden alle Werte auf einen ganzzahligen Wertebereich von 0 bis 2^p-1, $p \in \aleph$, transformiert. Dabei ist p für die realen Datensätze so gewählt, daß alle ungleichen Werte der Ursprungsdaten auch ungleich nach der Transformation sind und p minimal ist.

Für die univariaten künstlichen Daten wurde p=15 bzw. p=20 gewählt, wobei die Anzahl der Tupel bei den künstlichen Daten immer 100.000 beträgt. So wurden Datensätze mit kleiner ([0,...,32.767] mit p=15) und großer Domäne ([0,...,1.048.575] mit p=20) gebildet. Für die univariaten Datensätze, die der Normalverteilung folgen, wurden die Daten derart auf einen ganzzahligen Bereich abgebildet, daß ihr Mittelwert in der Mitte der Domäne liegt. Die Daten wurden so transformiert, daß alle erzeugten Datensätze innerhalb der Domäne liegen. Ebenso wurde bei den univariaten exponentialverteilten Daten verfahren.

Im bivariaten Fall folgen die Randverteilungen der künstlichen Datensätze den Verteilungen im univariaten Fall, sind aber nahezu unabhängig. Es wurde hier nur p = 15 betrachtet. Zudem wurden noch ein bivariater Datensatz mit stark korrelierten Daten (uk(15x15), s. Anhang) erzeugt. Die realen Datensätze lagen bereits bivariat vor (Tiger-Daten).

Mit Ausnahme der Selektivitätsschätzung durch Abbildung mehrdimensionaler Daten auf eindimensionale Daten anhand raumfüllender Kurven, lassen sich alle vorgestellten Selektivitätsschätzer auch direkt auf beliebigen reellwertigen Daten ausführen.

Die Kardinalität $|T|$ einer Testdatenmenge T wird entsprechend der Kardinalität einer Datenbankinstanz mit N bezeichnet. Da die Kardinalität N für einen bestimmten Testdatensatz fest ist, wird sie nur einmal erwähnt und ansonsten nicht mit aufgeführt.

5.3.2 Stichprobenmengen

Von allen erzeugten Datensätzen T wurden einfache Zufallsstichproben $S(T,n)$ gezogen mit jeweils n = 2.000 Werten. Dazu wurde als Stichprobenverfahren ohne Zurücklegen der Algorithmus von Vitter [Vitter 85] verwendet. Dieses Vorgehen und die Wahl der Stichprobengröße

stimmt mit dem in der Literatur über Selektivitätsschätzung in Datenbanksystemen verwendeten überein, vgl. z.B. [Poosala et al. 96].

Zusätzliche Experimente haben gezeigt, daß größere Stichprobenmengen auch zu besseren Schätzergebnissen führen, was die theoretischen Überlegungen aus Kapitel 4 bestätigt. Dieser Aspekt wird in Abschnitt 5.4.1 näher beleuchtet.

Im folgenden wird die Stichprobengröße wird meistens nicht weiter angegeben; in diesen Fällen ist die Stichprobengröße immer $n = 2000$.

5.3.3 Anfragen

Für jeden Test-Datensatz wurden vier Anfragemengen $A(m,q)$ generiert mit jeweils $m = 1000$ Bereichsanfragen der festen Größe $q = 1\%, 2\%, 5\%$ und 10% der zugrundeliegenden Domänengröße. Die Mittelpunkte der Anfragemengen wurden als Stichprobe aus dem Test-Datensatz gezogen. Es wurden dabei nur Anfragen akzeptiert, die ganz in der Domäne liegen. Dem beschriebenen Vorgehen liegt die Annahme zugrunde, daß die Verteilung der Anfragen der Verteilung der Daten entspricht. In Bereichen, wo viele Daten existieren, gibt es in realen Systemen sicherlich häufiger Anfragen, als in Bereichen, wo keine Daten existieren. Der Benutzer "weiß" i.a. wo seine Daten ungefähr liegen.

Die Anfragemengen in den Experimenten unterscheiden sich insofern von anderen Autoren (siehe z.B. [Poosala et al. 96]), als hier die Anfragegröße fest ist. So läßt sich der Einfluß der Anfragegröße auf die Güte der Selektivitätsschätzung beurteilen. Dies wird in Kapitel 5.4.2 betrachtet.

Ein weiterer wesentlicher Unterschied zu [Poosala et al. 96] ist, daß in der vorliegenden Arbeit explizit Anfragen der Form $Q(a,b)$, i.a. $a > 0$ bei linkem Rand 0, generiert wurden. Die in den Experimenten von [Poosala et al. 96] ausschließlich gewählten Anfragen der Form $Q(0,x)$ können abhängig von der wahren Dichte zu einer verfälschten Beurteilung der Ergebnisse führen. Einer der Autoren begründet die Wahl der Anfragen an anderer Stelle [Poosala 97] damit, daß gelte $\sigma(a, b) = \sigma(0, b) - \sigma(0, a)$. Das entsprechende gilt jedoch i.a. weder für die Schätzung der Selektivität

$$\text{i.a. } \hat{\sigma}(a, b) \neq \hat{\sigma}(0, b) - \hat{\sigma}(0, a) \tag{5.1}$$

noch für den (mittleren) absoluten oder relativen Fehler.

Angenommen, es gelte doch $\hat{\sigma}(a, b) = \hat{\sigma}(0, b) - \hat{\sigma}(0, a)$. Dann ist aber schon für den absoluten Selektivitätsfehler

$$\eta(\hat{\sigma}(a, b)) = |\sigma(a, b) - \hat{\sigma}(a, b)N|$$
$$= |(\sigma(0, b) - \sigma(0, a)) - (\hat{\sigma}(0, b) - \hat{\sigma}(0, a))N|$$

$$= |(\sigma(0,b) - \hat{\sigma}(0,b)N) - (\sigma(0,a) - \hat{\sigma}(0,a)N)|$$

$$\geq ||\sigma(0,b) - \hat{\sigma}(0,b)N| - |\sigma(0,a) - \hat{\sigma}(0,a)N|| = |\eta(\hat{\sigma}(0,b)) - \eta(\hat{\sigma}(0,a))|$$

wobei i.a. die echte Ungleichheit gilt. Die Gleichheit gilt in den Fällen, in denen $(\sigma(0,b) - \hat{\sigma}(0,b)N)$ und $(\sigma(0,a) - \hat{\sigma}(0,a)N)$ das gleiche Vorzeichen haben oder mindestens einer der beiden Terme gleich 0 ist. Ersteres tritt dann ein, wenn sowohl bei $Q(0,a)$ als auch $Q(0,b)$ gleichzeitig entweder überschätzt oder unterschätzt wurde. Dies ist i.a. nicht der Fall.

Dieser Unterschied kann noch viel deutlicher für den relativen Selektivitätsfehler ausfallen, wie man sich an einem einfachen konstruierten Beispiel (z.B. bei einer zugrundeliegenden Exponentialverteilung) mit einem beliebigen Selektivitätsschätzer verdeutlichen kann. Sei nun eine Anfrage $Q(a,b)$ gegeben mit a und b sehr groß, so daß sich z.B. eine korrekte Selektivität $\sigma(a,b) = 5$ und eine geschätzte Selektivität $\hat{\sigma}(a,b) = 6$ ergeben mögen. Seien die korrekten Selektivitäten $\sigma(0,b) = 1005$ und $\sigma(0,a) = 1000$. Die Schätzungen mögen ergeben $\hat{\sigma}(0,b) = 1007$ und $\hat{\sigma}(0,a) = 1001$. Dies führt zu einem relativen Fehler von $2/1500 \approx 0{,}2\%$ für $\hat{\sigma}(0,b)$ und $1/1000 = 0{,}1\%$ für $\hat{\sigma}(0,a)$, während der eigentliche Selektivitätsfehler für $\hat{\sigma}(a,b)$ $1/5 = 20\%$ beträgt. Dies ist ein Faktor von 100 bzw. 200. Ähnliche Konstellationen können bei den von [Poosala et al. 96] verwendeten Zipf-Verteilungen für große x auftreten.

Da die Anzahl der Anfragen in den folgenden Experimenten immer fest auf $m = 1000$ gesetzt ist, wird sie im folgenden nicht weiter explizit aufgeführt.

5.3.4 Verwendete Testdatenmengen

Ein Überblick über die verwendeten ein- und zweidimensionalen Datensätze ist in den folgenden Tabellen gegeben.

Univariate Testdaten

Tabelle 5.4 zeigt die Datensätze für die univariate Selektivitätsschätzung. Dabei wurde teilweise zwischen Verteilungen mit niedrigerer oder höherer Häufigkeit der Einzelwerte unterschieden (u(20)/u(15), n(20)/n(15), e(20)/e(15), rr(22)/rr(12)).

Datenmenge bzw. Verteilung	Bezeichnung	Domäne	Anzahl Tupel
Gleichverteilung, kleine Domäne	u(15)	$0..2^{15}-1$	100.000
Gleichverteilung, große Domäne	u(20)	$0..2^{20}-1$	100.000
Normalverteilung, kleine Domäne	n(15)	$0..2^{15}-1$	100.000

Tabelle 5.4: Eindimensionale Testdaten

Normalverteilung, große Domäne	n(20)	$0..2^{20}-1$	100.000
Exponentialverteilung, kleine Domäne	e(15)	$0..2^{15}-1$	100.000
Exponentialverteilung, große Domäne	e(20)	$0..2^{20}-1$	100.000
Los Angeles	la(16)	$0..2^{16}-1$	28.136
Arapahoe, 1. Komponente	ar1(21)	$0..2^{21}-1$	52.120
Arapahoe, 2. Komponente	ar2(18)	$0..2^{18}-1$	52.120
Rail road, 1. Komp., kleine Domäne	rr1(22)	$0..2^{22}-1$	257.942
Rail road, 2. Komp., kleine Domäne	rr2(22)	$0..2^{22}-1$	257.942
Rail road, 1. Komp., große Domäne	rr1(12)	$0..2^{12}-1$	257.942
Rail road, 2. Komp., große Domäne	rr2(12)	$0..2^{12}-1$	257.942

Tabelle 5.4: Eindimensionale Testdaten

Um den Hybridselektivitätsschätzer zu untersuchen wurden zusätzliche künstliche Datensätze mit künstlichen Sprungstellen erzeugt. Es handelt sich dabei um zwei Datensätze, bei denen zwischen den Sprungstellen eine Gleichverteilung (xu(*p*)) bzw. eine Exponentialverteilung (xe(*p*)) der Daten vorliegt. Diese Datensätze sind noch einmal in der Tabelle 5.5 separat aufgeführt:

Datenmenge bzw. Verteilung	**Bezeichnung**	**Domäne**	**Anzahl Tupel**
Sprungstellen mit lokaler Gleichverteilung	xu(15)	$0..2^{15}-1$	100.000
Sprungstellen mit lokaler Exponentialverteilung	xe(15)	$0..2^{15}-1$	100.000

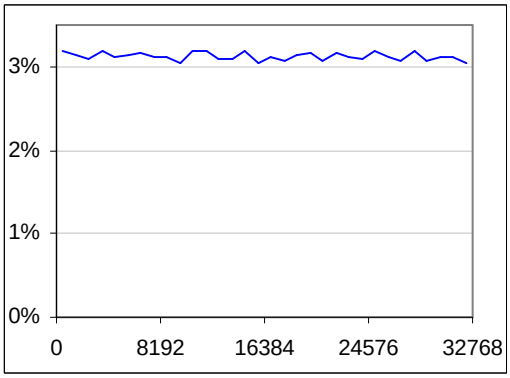
Tabelle 5.5: Eindimensionale Testdaten mit künstlichen Sprungstellen

Tabelle 5.6 listet die 8 jeweils zu dem Datensätzen xu(15) und xe(15) gehörenden Sprungstellen auf:

x	4.096	8.192	12.288	16.384	17.408	19.456	22.528	26.624

Tabelle 5.6: Sprungstellen der xu(15)- und xe(15)-Testdaten.

Um sich ein vorab ein grobes Bild von der Dichte der Testdatensätze zu machen, sind in den Abbildungen 5.1 - 5.7 Schätzungen der Dichtefunktion anhand von mit MS-Excel erzeugten Häufigkeitspolygonen dargestellt. Da es sich hierbei ebenfalls um (einfache) Schätzungen handelt, sind diese Abbildungen mit Vorsicht zu genießen. Dies gilt insbesondere im Hinblick auf Aussagen über die Glattheit bzw. analytische Stetigkeit der Kurven, die Anzahl der Modi oder das Aussehen der Dichte am Rand.



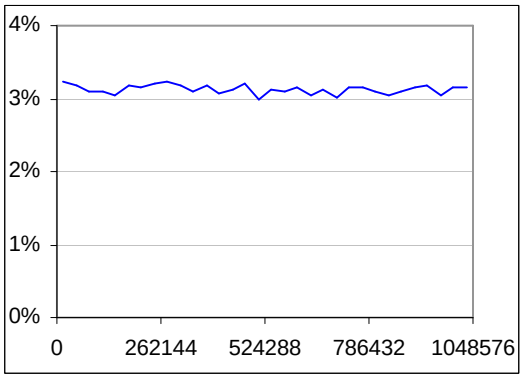
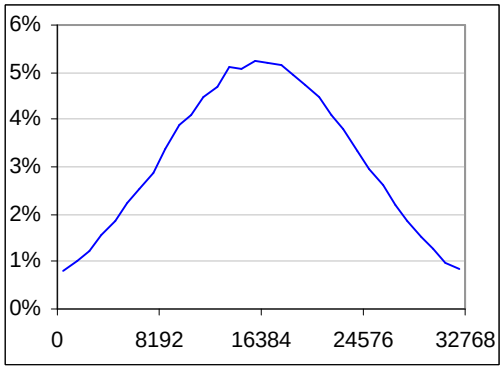


Abbildung 5.1: a) u(15)-Daten (HP mit 32 Bins). **b)** u(20)-Daten (HP mit 32 Bins).



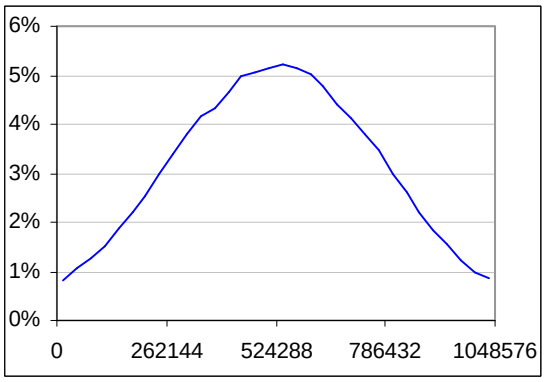
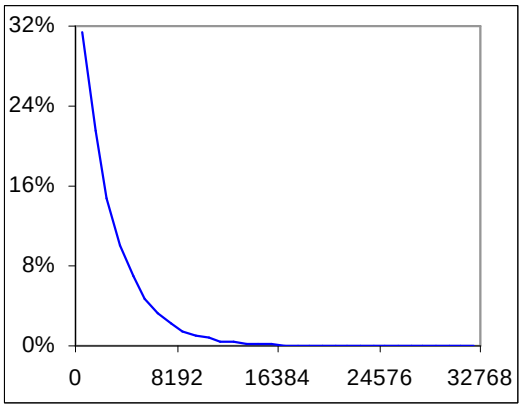


Abbildung 5.2: a) n(15)-Daten (HP mit 32 Bins). **b)** n(20)-Daten (HP mit 32 Bins).



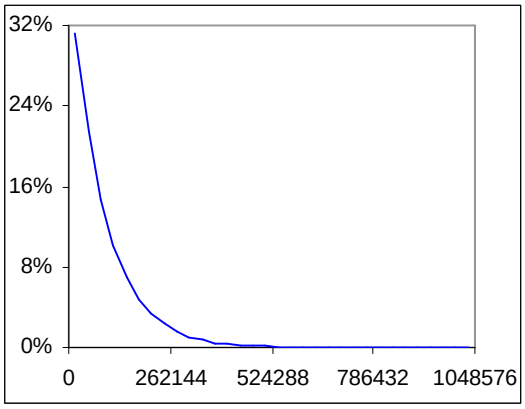


Abbildung 5.3: a) e(15)-Daten (HP mit 32 Bins). **b)** e(20)-Daten (HP mit 32 Bins).

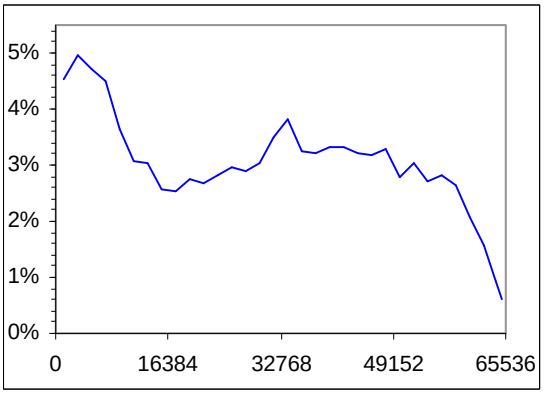
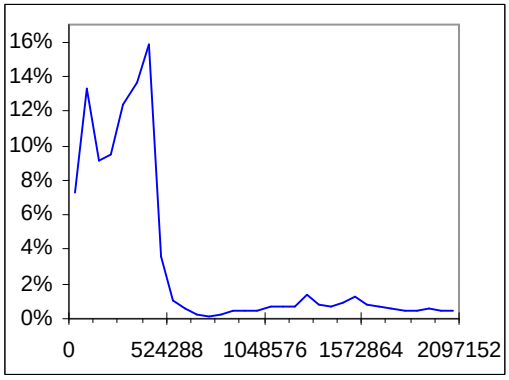


Abbildung 5.4: la(16)-Daten (HP mit 32 Bins).



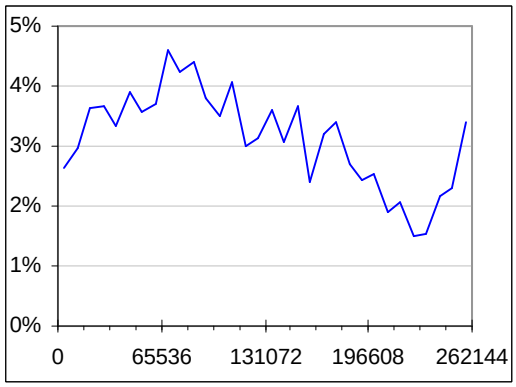
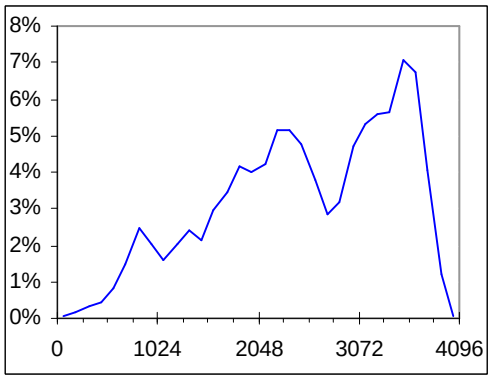


Abbildung 5.5: a) ar1(21)-Daten (HP mit 32 Bins). **b)** ar2(18)-Daten (HP mit 32 Bins).



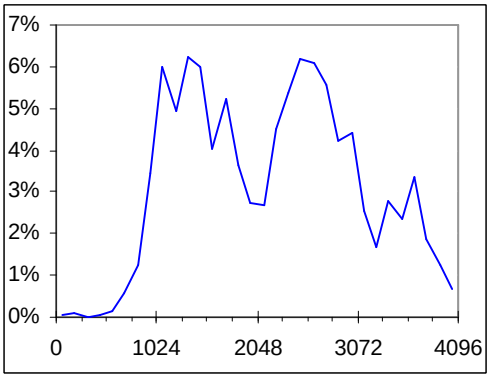


Abbildung 5.6: a) rr1(12)-Daten (HP mit 32 Bins).**b)** rr2(12)-Daten (HP mit 32 Bins).

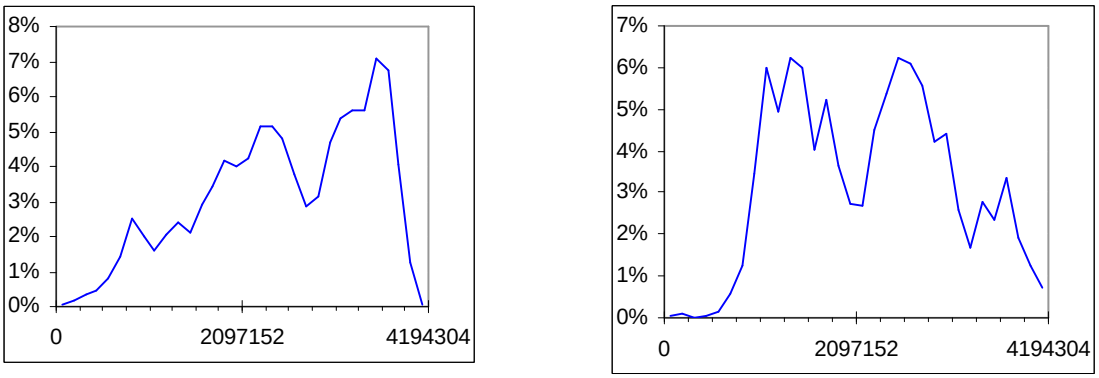


Abbildung 5.7: a) rr1(22)-Daten (HP mit 32 Bins).**b)** rr2(22)-Daten (HP mit 32 Bins).

In den folgenden Abbildungen 5.8 und 5.9 finden sich weiterhin grobe Dichteschätzungen der beiden speziellen Testdatensätze xu(15) und xe(15) für den Hybridselektivitätsschätzer.

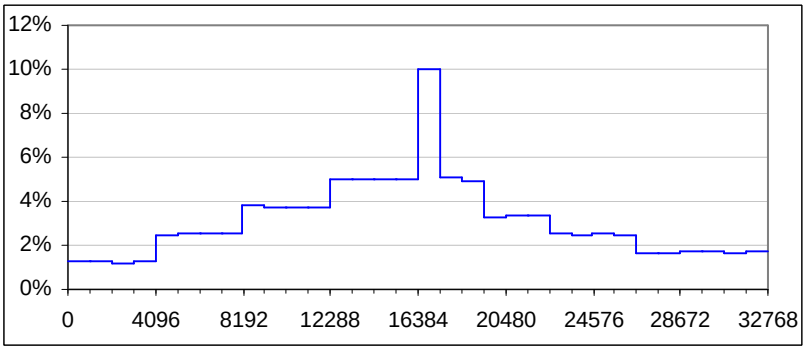


Abbildung 5.8: xu(15)-Daten durch Histogrammschätzer mit Binweite h = 1.024 visualisiert.

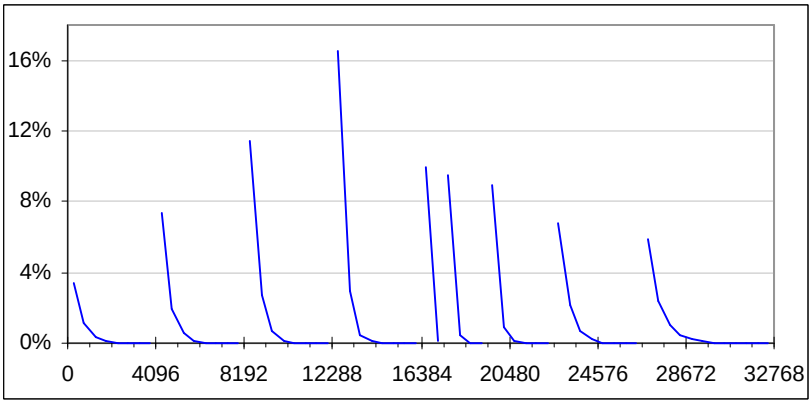


Abbildung 5.9: xe(15)-Daten durch Häufigkeitspolygonschätzer mit Binweite h = 256 visualisiert. Die Verbindungen an den (bekannten) Sprungstellen sind ausgelassen worden.

Bivariate Testdaten

Tabelle 5.7 zeigt die Datensätze der bivariaten Selektivitätsschätzung. Es handelt sich im Prinzip um die gleichen Testdaten wie im univariaten Fall mit folgenden Abweichungen: Bei den künstlichen Testdaten erfolgte eine Beschränkung auf p = 15, die Randverteilung der Testdaten entspricht der Verteilung im univariaten Fall. Abgesehen von den uk(15x15)-Testdaten wurden die Randverteilungen unabhängig voneinander erzeugt. Die uk(15x15)-Testdaten sind stark korreliert, um den Einfluß der Korrelation auf die Schätzverfahren zu untersuchen, siehe Kapitel 5.6.5. Auch bei den realen Testdaten folgt die Randverteilung der Verteilung der univariaten Testdaten. Man beachte, daß bei den Arapahoe-Testdaten $p_1 \neq p_2$. Die Anzahl der Tupel entspricht den univariaten Testdaten.

Datenmenge bzw. Verteilung	Bezeichnung	Bereich 1. Dim.	Bereich 2. Dim.	Anzahl Tupel
Gleichverteilung	u(15x15)	$0..2^{15}$-1	$0..2^{15}$-1	100.000
Gleichverteilung korreliert	uk(15x15)	$0..2^{15}$-1	$0..2^{15}$-1	100.000
Normalverteilung	n(15x15)	$0..2^{15}$-1	$0..2^{15}$-1	100.000
Exponentialverteilung	e(15x15)	$0..2^{15}$-1	$0..2^{15}$-1	100.000
Los Angeles	la(16x16)	$0..2^{16}$-1	$0..2^{16}$-1	28.136
Arapahoe County	ar(21x18)	$0..2^{21}$-1	$0..2^{18}$-1	52.120
Rail road (kleinere Dichte)	rr(22x22)	$0..2^{22}$-1	$0..2^{22}$-1	257.942
Rail road (höhere Dichte)	rr(12x12)	$0..2^{12}$-1	$0..2^{12}$-1	257.942

Tabelle 5.7: Zweidimensionale Daten

Um sich auch hier vorab ein grobes Bild von der Dichte der Daten machen zu können sind in den Abbildungen 5.10 - 5.14 die Scatter-plots der zweidimensionalen Daten dargestellt. Dabei wurde jeweils eine Stichprobe der Größe 2000 zugrunde gelegt.

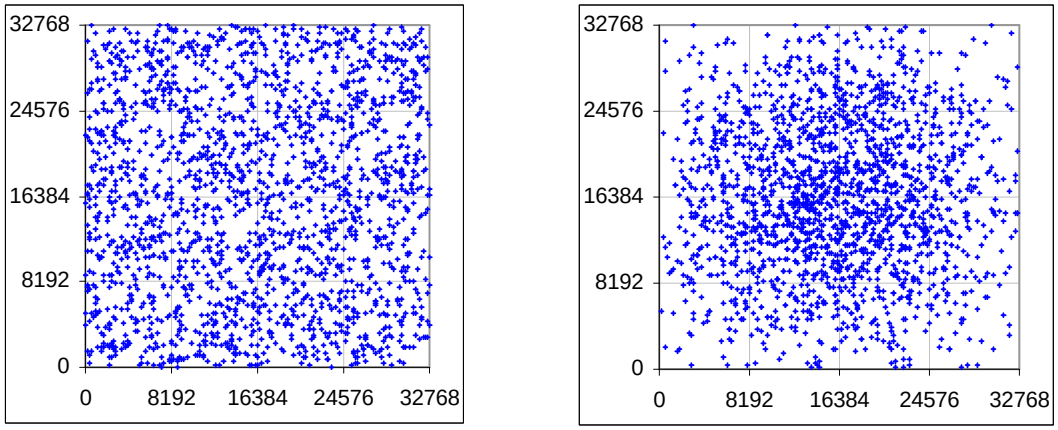


Abbildung 5.10: a) Scatterplot der u(15x15)-Daten. **b)** Scatterplot der n(15x15)-Daten.

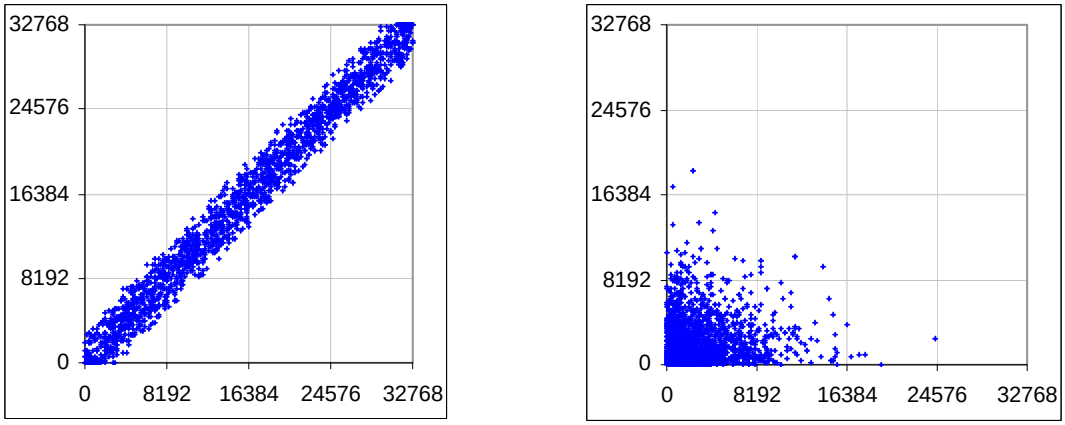


Abbildung 5.11: a) Scatterplot der uk(15x15)-Daten. **b)** Scatterplot der e(15x15)-Daten.

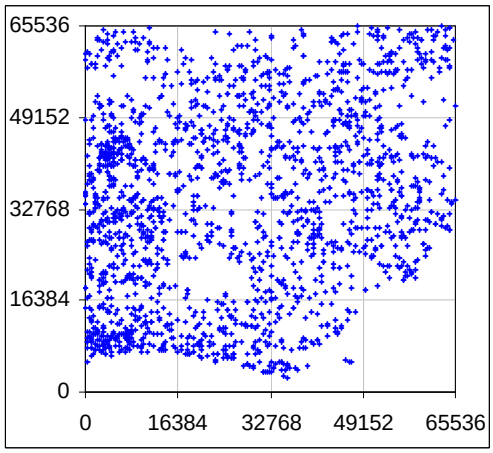


Abbildung 5.12: Scatterplot der la(16x16)-Daten bei einer Stichprobe der Größe n = 2.000.

Ein Scatterplot der Grundgesamtheit der LA-Testdaten findet sich im Anhang, Kapitel A.5.

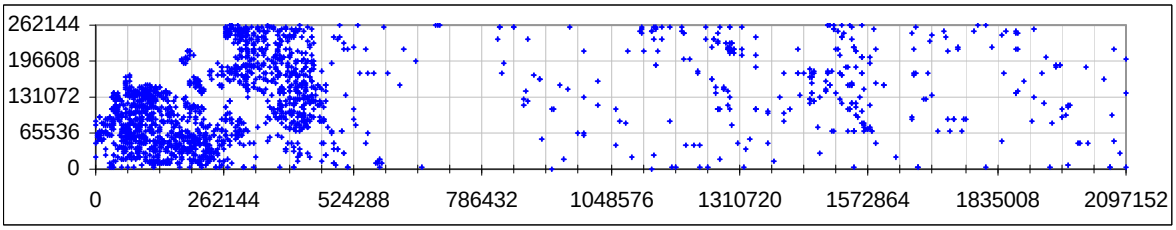
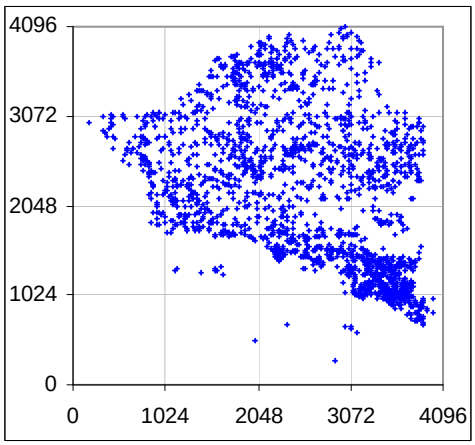


Abbildung 5.13: Scatterplot der ar(21x18)-Daten bei einer Stichprobe der Größe n = 2.000.

Ein Scatterplot einer Stichprobe der Größe n = 30.000 der Arapahoe-Testdaten findet sich im Anhang, Kapitel A.5.



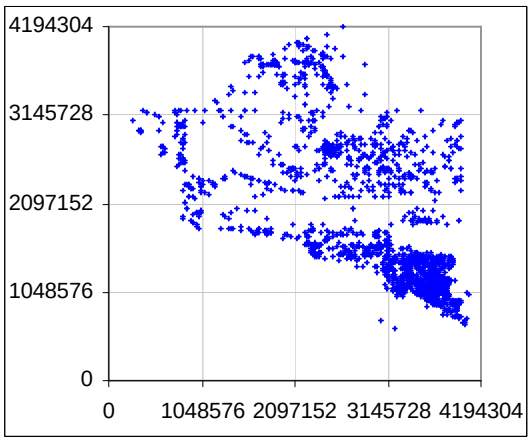


Abbildung 5.14: a) Scatterplot der rr(12x12)-Daten. **b)** Scatterplot der rr(22x22)-Daten.

5.4 Allgemeine Ergebnisse

Zunächst werden allgemeine Ergebnisse und Zusammenhänge aufgezeigt, die für alle verwendeten nicht-parametrischen Selektivitätsschätzer unabhängig von der Dimension der Daten gelten. Bereits aus Kapitel 4 ist der theoretische Einfluß der Stichprobengröße auf die Qualität der Schätzung bekannt. Die praktischen Ergebnisse bestätigen dies. Auch die Größe und Lage der Anfragen hat einen wesentlichen Einfluß auf den Fehler der Schätzung. Diese Zusammenhänge werden im folgenden kurz geschildert und beispielhaft anhand von Experimenten belegt. Zuletzt werden noch Bemerkungen über den Einfluß der wahren Dichte auf die Schätzung angegeben.

Der Bedeutung des Bandbreitenparameters für die Güte der Selektivitätsschätzung wurde bereits in den vorigen Kapiteln hervorgehoben. Die dort beschriebenen Ergebnisse hängen von der Dimension und dem verwendeten Schätzverfahren ab und werden daher ausführlich in den Kapiteln 5.5 und 5.6 validiert.

5.4.1 Der Einfluß der Stichprobengröße

Zusätzliche Experimente haben gezeigt, daß größere Stichprobenmengen auch zu besseren Schätzergebnissen führen, was die theoretischen Überlegungen aus Kapitel 4 bestätigt. Abbildung 5.15 zeigt die Abhängigkeit des mittleren relativen Selektivitätsfehlers MRSF von der Stichprobengröße *n* bei verschiedenen Selektivitätsschätzern für die n(20) Testdaten und einer festen Anfragegröße *q* = 1% des Datenraums. Der MRSF ist nur teilweise dargestellt, damit der Unterschied zwischen den anderen Kurven bei gröberer Skalierung der Grafik nicht verschwindet. Bei allen drei Selektivitätsschätzern kann die Qualität der Schätzung durch größere Stichprobenmengen i.a. erhöht werden, jedoch müssen der höhere Berechnungsaufwand und die größeren Ein-/Ausgabekosten berücksichtigt werden.

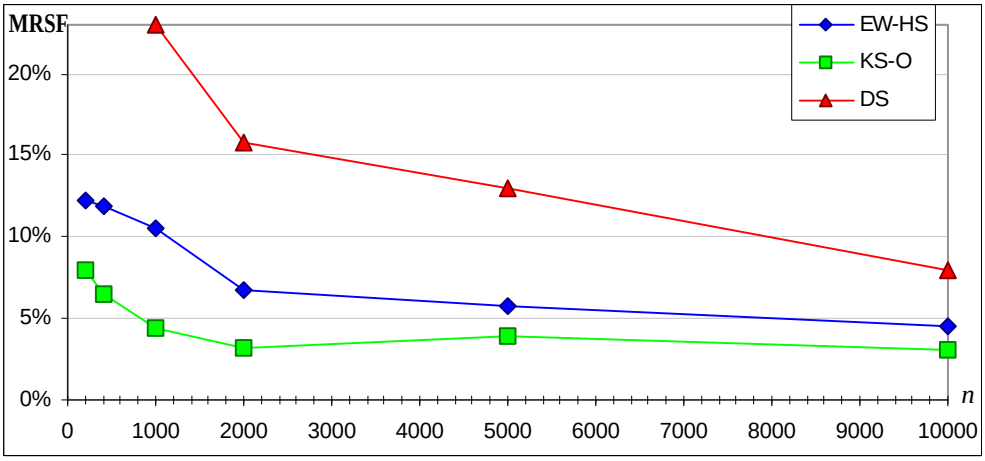


Abbildung 5.15: MRSF bei n(20) Daten mit *q* = 1% abhängig von *n* beim DS, EW-HS und KS-O.

5.4.2 Der Einfluß der Anfragegröße

Des weiteren sei der Einfluß der Anfragegröße *q* auf den mittleren relativen Selektivitätsfehler diskutiert. Abbildung 5.16 zeigt die Ergebnisse für den univariaten Equi-Width Histogrammschätzer bei verschiedenen Testdatensätzen und mit vier verschiedenen Anfragegrößen (1%, 2%, 5% und 10% der Domänengröße). Ähnliche Resultate wurden mit den anderen Selektivitätsschätzern erzielt. Aus der Abbildung wird ersichtlich, daß der mittlere relative Selektivitätsfehler bei kleineren Anfragen größer ist als bei größeren Anfragen. Der Schwerpunkt der Arbeit liegt auf den realistischeren kleineren Anfragen, so daß im folgenden nur noch die Resultate bei 1%-Anfragen berichtet werden.

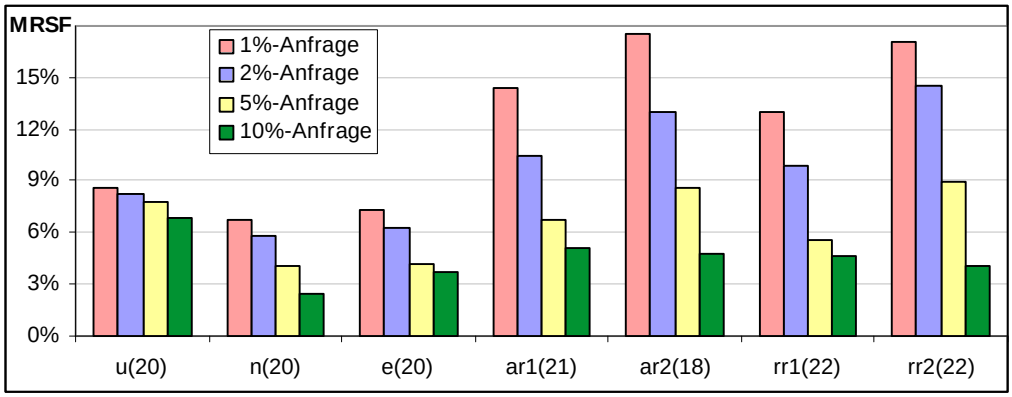


Abbildung 5.16: MRSF des HS bei Anfragemengen mit Anfragen unterschiedlicher Größe *q*

5.4.3 Der Einfluß der wahren Dichte

Bei der Gleichverteilung (u(*p*)-Daten) handelt es sich um einen Spezialfall, bei dem die verschiedenen nicht-parametrischen Schätzer im Vergleich zueinander zu unüblichen Ergebnissen kommen. So ist die Gleichverteilungsannahme hier (nahezu) optimal und liefert einen der geringsten Fehler (u(20): MRSF = 2,89%, u(15): MRSF = 2,39%). Dies ist bei den anderen Dichteverteilungen nicht der Fall, dort ist der relative Selektivitätsfehler der Gleichverteilungsannahme i.a. extrem hoch. Die Gleichverteilungsannahme entspricht einem Histogrammschätzer mit lediglich einem Bin. Von daher unterscheiden sich die Ergebnisse des Histogrammselektivitätsschätzers mit optimaler gefundener Bandbreite OGB (u(20): MRSF = 2,87%, u(15): MRSF = 2,39%) und Gleichverteilungsannahme kaum, allerdings liefern die Verfahren zur Schätzung der AOB i.a. eine höhere Anzahl von Bins (u(20): 13 Bins, u(15): 13 Bins), so daß hier das Ergebnis schlechter ausfällt (u(20): MRSF = 8,5%, u(15): MRSF = 5,6%). Für den Kernselektivitätsschätzer wäre eine Bandbreite erforderlich, die die gesamte Domäne abdeckt. Von daher werden die Ergebnisse des Kernselektivitätsschätzers (mit Randbehandlung) mit wachsender Größe der Bandbreite besser (u(20): MRSF= 2,85% bei h = 400.000, u(15): MRSF = 2,80% bei h = 10.000). Leider schätzen die Verfahren zur Bandbreitenbestimmung die AOB zu gering (u(20): MRSF = 4,0% bei h = 155.527, u(15): MRSF = 4,2% bei h = 4.853). Da die gleichverteilten Daten relativ viel Masse am Rand der Domäne besitzen, ist hier bei der Kernselektivitätsschätzung eine Randbehandlung erforderlich.

Die folgenden Experimente haben teilweise drastische Unterschiede in den Ergebnissen bei realen und künstlichen Testdaten gezeigt. Die Vermutung, daß es sich dabei um Unstetigkeiten in der wahren Dichte handelt, wird durch die Ergebnisse beim univariaten Hybridselektivitätsschätzer in Abschnitt 5.5.6 bestätigt. Deren Einfluß kann so drastisch sein, daß Schätzverfahren mit einem besseren AMISE zu schlechteren Ergebnissen führen, da die in der Theorie angenommenen Bedingungen nicht erfüllt sind. Im univariaten Fall wird dem Einfluß von Unstetigkeiten durch die Identifikation von “Sprungstellen” und Anwendung des Hybridselektivitätsschätzers geeignet begegnet.

Auch die Korrelation der Daten im multivariaten Fall kann einen starken negativen Einfluß auf die Ergebnisse der Selektivitätsschätzung haben. Dies gilt insbesondere für die verschiedenartigen Histogramm- und Kernselektivitätsschätzer, da hierdurch die Voraussetzung der Unabhängigkeit der Variablen nicht erfüllt ist. Auf Experimente mit korrelierten Testdaten wird in Abschnitt 5.6.5 im bivariaten Fall eingegangen. Die Experimente bestätigen zudem, daß sich durch Schätzung der asymptotisch optimalen Bandbreite unter Berücksichtigung des Korellationskoeffizienten (vgl. Abschnitt 4.4.5) dem negativen Einfluß der Korrelation geeignet begegnen läßt.

5.5 Ergebnisse im univariaten Fall

In diesem Kapitel werden zunächst die Ergebnisse der univariaten Selektivitätsschätzung präsentiert. Dazu werden zum einen die verschiedenen Histogrammselektivitätsschätzer (Equi-Width, Equi-Depth und Max-Diff) miteinander verglichen, zum anderen die Kernselektivitätsschätzer mit und ohne Randbetrachtung. Als interessante Alternative werden die Ergebnisse weiterhin mit dem ASH-Selektivitätsschätzer verglichen. Ein besonderes Augenmerk wird auf die Bestimmung der asymptotisch optimalen Bandbreite (AOB) gelegt. Zuletzt werden die Ergebnisse des Hybridselektivitätsschätzers als das für reale Datensätze geeignetere Verfahren vorgestellt.

Die Ergebnisse der Selektivitätsschätzung mittels der Gleichverteilungsannahme (GS) und des direkten Selektivitätsschätzers (DS) sollen aufgrund der Einfachheit und weiten Verbreitung der Verfahren als Referenzwerte in den folgenden Vergleichen dienen. Deshalb werden zunächst die Ergebnisse mit diesen beiden Verfahren im folgenden kurz vorgestellt und diskutiert. Die Verfahren wurden jeweils basierend auf einer Stichprobe von 2000 Elementen mit 1000 Testdaten der Größe 1%, 2%, 5% und 10% getestet.

5.5.1 Ergebnisse der Selektivitätsschätzung mittels Gleichverteilungsannahme

Da es sich bei der Gleichverteilungsannahme um ein parametrisches Verfahren handelt, sind gute Ergebnisse nur zu erwarten, wenn die wahre Verteilung der Daten näherungsweise zur Familie der Gleichverteilungen gehört. Dies ist bei den Testdatensätzen lediglich bei den gleichverteilten Datensätzen u(*p*) der Fall, wo der mittlere relative Selektivitätsfehler (MRSF) zwischen 0,4% (u(15), 10%-Anfragen) und 2,9% (u(20), 1%-Anfragen) liegt. Bei allen anderen Testdatensätzen sind die Ergebnisse nicht akzeptabel, insbesondere bei schiefen Verteilungen wie z.B. den exponential verteilten Datensätzen e(*p*), wo der MRSF sogar auf über 200% (e(15), 1%-Anfragen) ansteigen kann. Abbildung 5.17 zeigt die unterschiedlichen Ergebnisse der GS bei den verschiedenen Verteilungen und einer Anfragegröße von 1%.

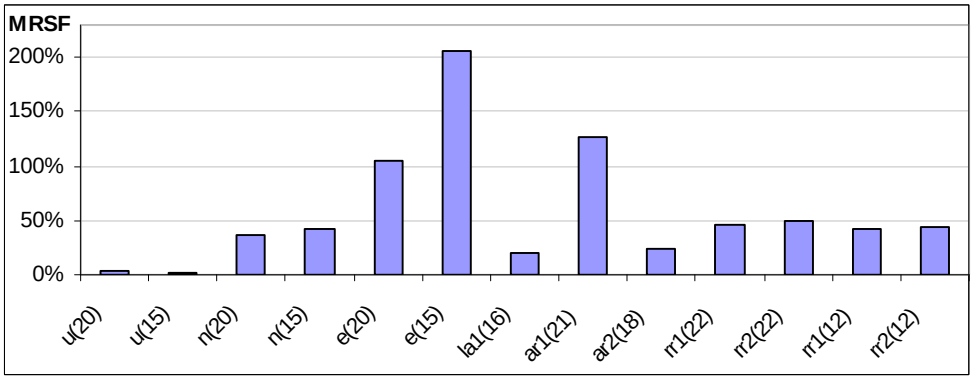


Abbildung 5.17: MRSF der Selektivitätsschätzung aufgrund der Gleichverteilungsannahme (GS) bei 1%-Anfragen.

5.5.2 Ergebnisse der direkten Selektivitätsschätzung

Die Ergebnisse der direkten Selektivitätsschätzung sind unabhängig von der zugrunde liegenden wahren Verteilung und zunächst als durchaus akzeptabel zu bezeichnen. Abbildung 5.18 zeigt die unterschiedlichen Ergebnisse der DS bei den verschiedenen Verteilungen und einer Anfragegröße von 1%. Der MRSF liegt hier zwischen 9,7% (e(20)) und 18,5% (ar2(18)).

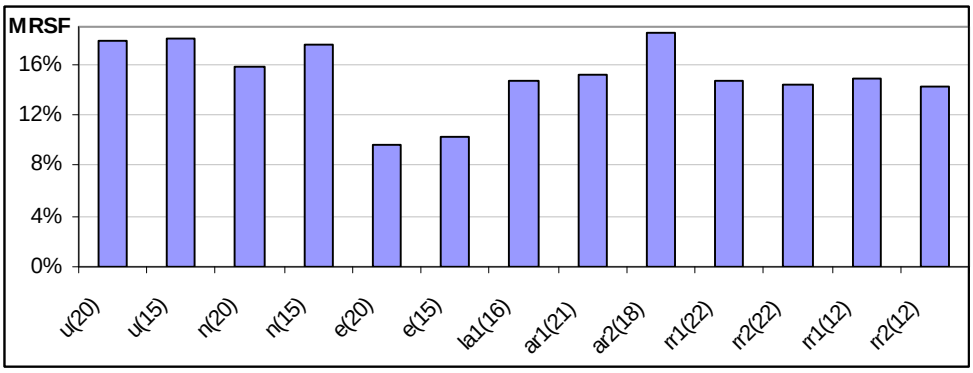


Abbildung 5.18: MRSF der direkten Selektivitätsschätzung (DS) bei 1%-Anfragen.

Es ist daher Ziel der folgenden Experimente insbesondere die Überlegenheit der verfeinerten nicht-parametrischen Verfahren gegenüber der direkten Selektivitätsschätzung aufzuzeigen.

5.5.3 Ergebnisse der verschiedenen Histogrammselektivitätsschätzer

Im folgenden werden zunächst die Ergebnisse der verschiedenen Histogrammselektivitätsschätzer (HS) miteinander verglichen. Dazu wurden aus den Stichproben Equi-Depth (ED), Equi-Width (EW) und Max-Diff (MD) Histogramme mit unterschiedlicher Anzahl Bins erzeugt

und mit jeweils 1000 Anfragen der Größen 1%, 2%, 5% und 10% getestet. Insbesondere soll der Einfluß der Binweite (bzw. äquivalent der Anzahl der Bins) auf die Güte der Schätzung herausgestellt werden.

Zuerst wurden die verschiedenen Histogrammselektivitätsschätzer mit variierender Anzahl von Bins getestet und die jeweils besten Ergebnisse miteinander verglichen. Man beachte, daß dabei für die verschiedenen Histogrammselektivitätsschätzer auch bei gleichen Testdaten unterschiedliche Binweiten bzw. eine unterschiedliche Anzahl von Bins vorliegen können.

Da die Gleichverteilungsannahme aus vorigem Abschnitt einem Histogrammschätzer mit einem einzigen Bin entspricht, ist bei den gleichverteilten Testdaten zu erwarten, daß die optimale Anzahl von Bins in diesem Falle ebenfalls - unabhängig vom jeweiligen Partitionierungsverfahren - einem einzigen Bin entspricht. Dies wird durch die Experimente weitestgehend bestätigt, wobei in einigen Fällen der HS mit zwei oder drei Bins minimalst bessere Ergebnisse aufweisen kann. Z.B. ergaben die Experimente bei den u(20)-Testdaten und 1%-Anfragen bei einem Bin einen MRSF von 2,89% während der EW-HS bei drei Bins sogar nur einen MRSF von 2,87% aufwies. Dies läßt sich jedoch durchaus durch Artefakte in der zugrundeliegenden Stichprobe erklären. Es zeigt sich, daß die Ergebnisse beim Equi-Width-HS in den meisten Fällen am besten sind. Nur in wenigen Fällen (la1(16), ar2(18)) ist der Equi-Depth-HS besser. Weniger gut schneidet der Max-Diff-HS ab, der abgesehen von den u(*p*)-Testdaten schlechtere Ergebnisse aufweist. Auffällig ist beim Max-Diff-HS auch, daß hier eine hohe Anzahl von Bins erforderlich ist, um das optimale Ergebnis zu erhalten. Die hier erzielten Ergebnisse weichen von in anderen Veröffentlichungen präsentierten Ergebnissen ab. Z.B. berichten [Poosala et al. 96], daß das Max-Diff-Verfahren in ihren Ergebnissen besser abschneidet. Der Grund mag in den unterschiedlichen Modellannahmen liegen (siehe Vorwort). In Abbildung 5.19 ist der mittlere relative Selektivitätsfehler MRSF der verschiedenen HS für die Testfälle bei 1%-Anfragen graphisch dargestellt.

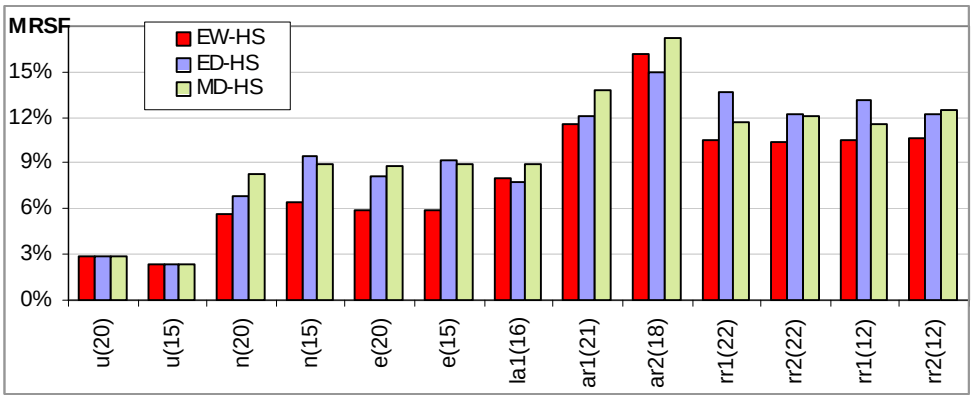


Abbildung 5.19: MRSF des HS bei verschiedenen Histogrammtypen mit jeweils OGB bei 1%-Anfragen.

Abbildung 5.20 zeigt für einige Testdaten den MRSF der verschiedenen HS bei 1%-Anfragen mit laufender Binweite. Für die EW-HS ist zum Vergleich die mittels Normalskalierungsregel

geschätzte asymptotisch optimale Binweite (AOB) angegeben. Dabei wurde das Maximum der beiden Varianten - Schätzung mittels Standardabweichung bzw. mittels Interquartilsabstand - gewählt. Typisch für die Grafiken ist, daß die Kurven für EW-HS und ED-HS einen ähnlichen Verlauf zeigen, während der MD-HS erst bei einer sehr viel höheren Anzahl von Bins sein Minimum findet. Weiterhin ist zu beachten, daß die Kurven lokal stark oszillieren. Dies bedeutet, daß z.B. ein Bin mehr oder weniger einen größeren MRSF erzeugen kann als z.B. zwei oder drei Bins mehr oder weniger. Der Grund liegt unter anderem in der Abhängigkeit der HS vom Startwert. Z.B. ist für die rr1(22) Daten die optimale gefundene Anzahl von Bins 74 mit einem MRSF von 10,6%, wobei bei einer Anzahl von 38 Bins ein MRSF von 10,7% vorliegt, was nur ein geringer Unterschied ist. Dagegen ist bei 44 Bins ein weitaus höherer MRSF von 14,5% vor. Bei der mit der Normalskalierungsregel bestimmten Anzahl von 18 Bins liegt ein MRSF von 13,0% vor, der nur suboptimal ist. Insofern sind die Zahlen der optimalen gefundenen Anzahl Bins als Richtwerte zu verstehen. Die Schätzung einer asymptotisch optimalen Bandbreite hat daher eher tendenziellen Charakter. Eine weitaus kleinere Anzahl von Bins sollte ebenso vermieden werden wie eine zu große Anzahl Bins.

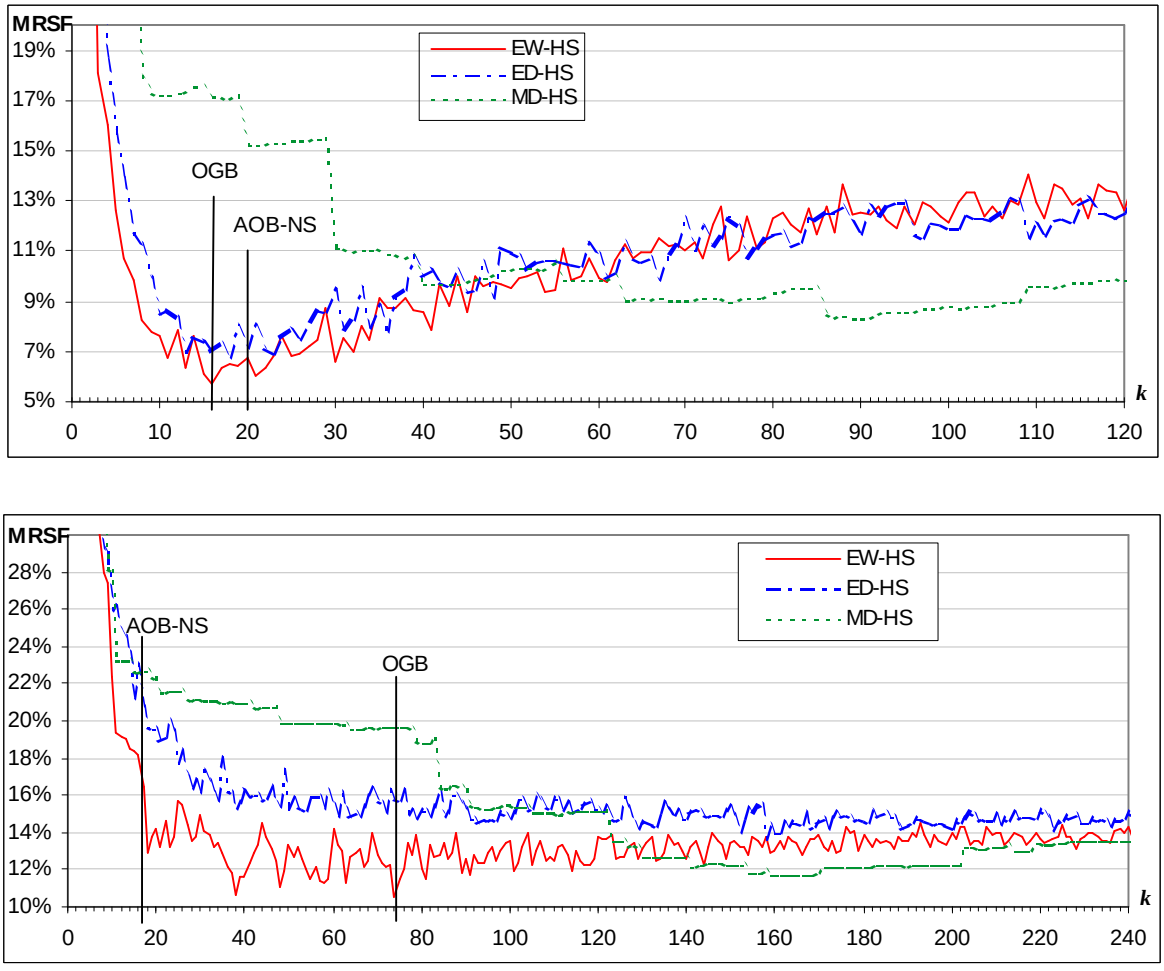


Abbildung 5.20: MRSF des HS bei 1%-Anfragen und verschiedenen Histogrammtypen mit variabler Binweite für die Testdatensätze n(20) (oben) und rr1(22) (unten).

Für den obigen normalverteilten Datensatz *n*(20) seien in Abbildung 5.21 die Histogramme für die verschiedenen Partitionierverfahren mit jeweils optimaler gefundenener Anzahl Bins miteinander verglichen. Dabei zeigt sich, daß das EW-Histogramm die Normalverteilung am besten darstellt. Auch das ED-Histogramm trifft die Normalverteilungskurve sehr gut, glättet aber etwas zu stark am Rand. Dagegen besitzt das MD-Histogramm am Rand zu viele Klassen, so daß hier Artefakte in der Stichprobe herausgestellt werden.

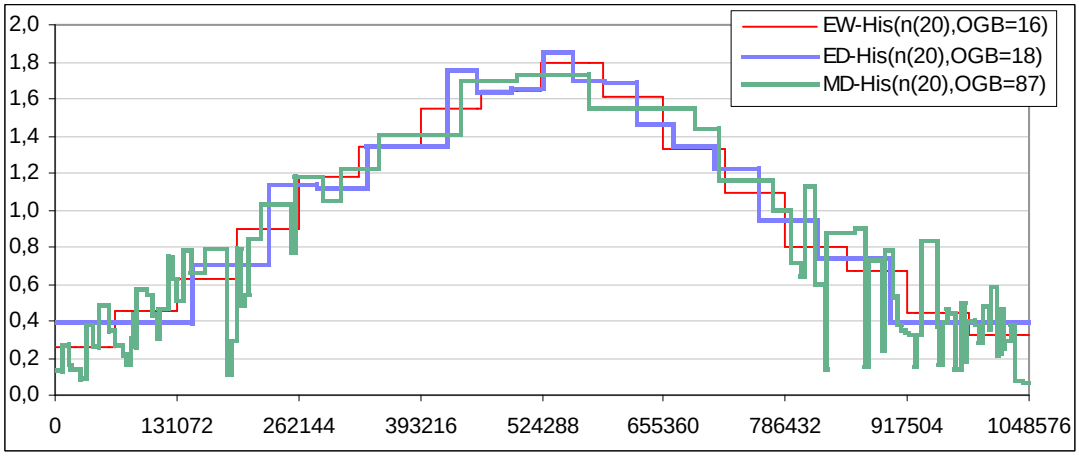


Abbildung 5.21: Histogramme bei verschiedenen Partitionierverfahren mit jeweils optimaler gefundener Anzahl Bins bei Datensatz n(20).

Anschließend sollen in Abbildung 5.22 für die Equi-Width HS die zugehörigen Histogramme zum einen mit optimaler gefundener Anzahl Bins (OGB) und zum anderen mit geschätzter asymptotisch optimaler Anzahl Bins (AOB) dargestellt werden. Die Schätzung der AOB mit der Normalskalierungsregel ergab auf der vorliegenden Stichprobe der normalverteilten Testdaten 20 Bins, dagegen brachte das Histogramm mit 16 Bins bei laufender Anzahl Bins den besten MRSF. Wie aus der Abbildung ersichtlich ist, sind die beiden Histogramme sehr ähnlich, was mit der geringen Abweichung des MRSF korrespondiert. Dagegen unterscheiden sich die Histogramme bei den realen rr1(22)-Testdaten deutlich. Die Schätzung der AOB mit der Normalskalierungsregel ergab auf der vorliegenden Stichprobe 17 Bins, dagegen brachte das Histogramm mit 74 Bins bei laufender Anzahl Bins den besten MRSF. Auch der MRSF unterscheidet sich deutlicher (10,6% zu 13,0%). Dies liegt daran, daß die Normalskalierungsregel bei nicht-normalverteilten Daten i.a. deutlich überglättet.

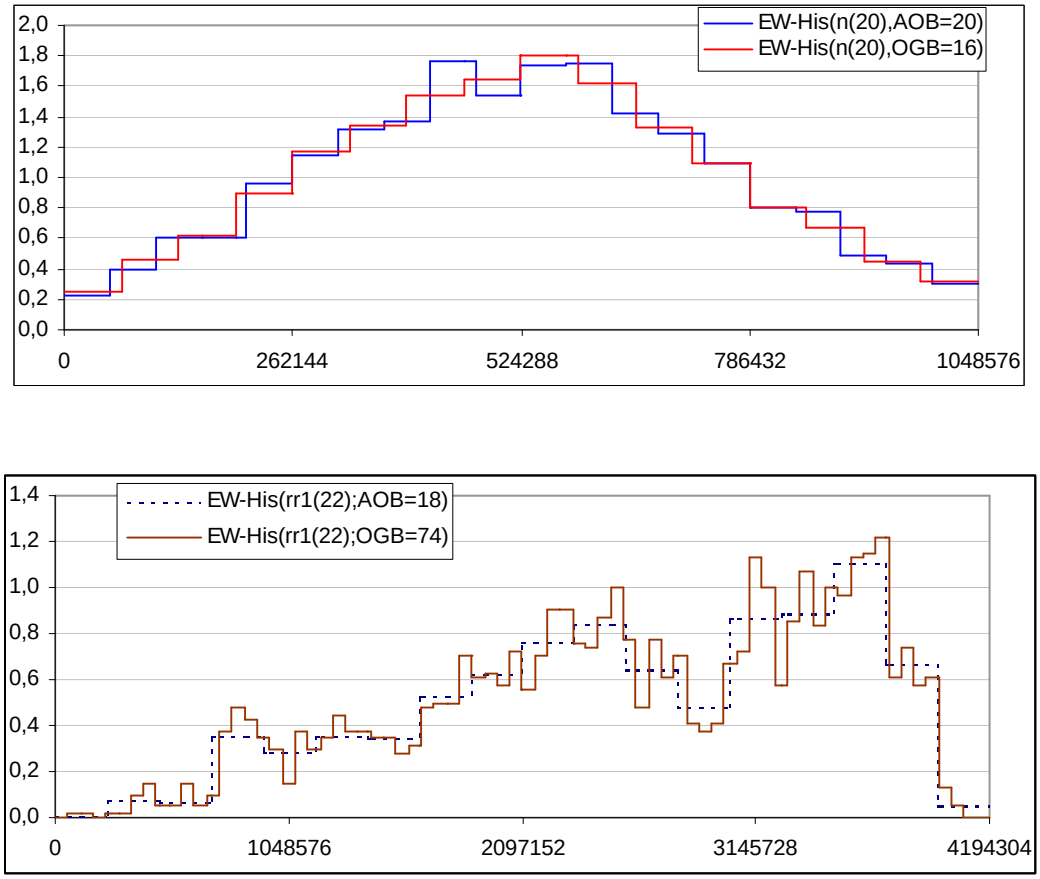


Abbildung 5.22: Histogramme des EW-HS mit optimaler gefundener (OGB) und geschätzter asymptotisch optimaler (AOB) Anzahl Bins bei n(20)- (oben) und rr1(22)- (unten) Testdaten.

In Tabelle 5.8 sind noch einmal für die verschiedenen Testdaten die Anzahl der Bins und der mittlere relative Selektivitätsfehler (MRSF) für den EW-HS einmal bei optimaler gefundener Anzahl Bins (OGB) und zum anderen bei der Anzahl Bins geschätzt mit der Normalskalierungsregel (AOB-NS) aufgeführt. Dies bestätigt die oben beschriebenen Beobachtungen.

Testdatensatz	OGB	AOB-NS	MRSF(OGB)	MRSF(AOB-NS)
u(20)	3	13	2,9	8,5
u(15)	1	13	2,4	5,6
n(20)	16	20	5,7	6,8
n(15)	15	20	6,4	8,4
e(20)	56	69	6,0	7,3
e(15)	61	69	5,8	7,7
la1(16)	31	13	8,0	9,0

Tabelle 5.8: Vergleich der Anzahl Bins und des MRSF bei 1%-Anfragen für den EW-HS mit OGB und AOB-NS.

Testdatensatz	OGB	AOB-NS	MRSF(OGB)	MRSF(AOB-NS)
ar1(18)	102	51	11,5	14,4
ar2(21)	94	14	16,1	17,6
rr1(22)	74	18	10,6	13,0
rr2(22)	51	19	10,3	17,1
rr1(12)	42	18	10,5	12,9
rr2(12)	51	18	10,7	16,7

Tabelle 5.8: Vergleich der Anzahl Bins und des MRSF bei 1%-Anfragen für den EW-HS mit OGB und AOB-NS.

Abbildung 5.23 vergleicht den Equi-Width Histogrammselektivitätsschätzer (EW-HS) einmal mit optimaler gefundener Binweite (OGB) und zum anderen mit geschätzter asymptotisch optimaler Binweite (AOB) mit dem direkten Selektivitätsschätzer (DS). Ein solcher Vergleich bzgl. den anderen HS (Equi-Depth, Max-Diff) ist nicht möglich, da hierfür keine Verfahren zur Bestimmung der AOB bekannt sind. Abbildung 5.23 zeigt zum einen wie erwartet, daß der EW-HS mit OGB noch durchaus bessere Ergebnisse zeigt als der EW-HS mit geschätzter AOB, wobei der Sonderstatus der gleichverteilten Testdaten berücksichtigt werden muß (s.o.). Bessere Schätzverfahren zur Bestimmung der AOB werden im Abschnitt 5.5.5 im Zusammenhang mit dem Kernselektivitätsschätzer diskutiert. Zum anderen zeigt sich die deutliche Überlegenheit der EW-HS gegenüber den DS selbst bei nur geschätzter AOB. Einzig bei den rr2(p) Daten ist der DS besser als der EW-HS(AOB), wobei dies nicht daran liegt, daß der DS bei diesen Testdaten so gut ist, sondern daran daß das Verfahren zur Bestimmung der AOB in diesem Falle eher versagt. Der EW-HS(OGB) ist in allen Testfällen besser als der DS.

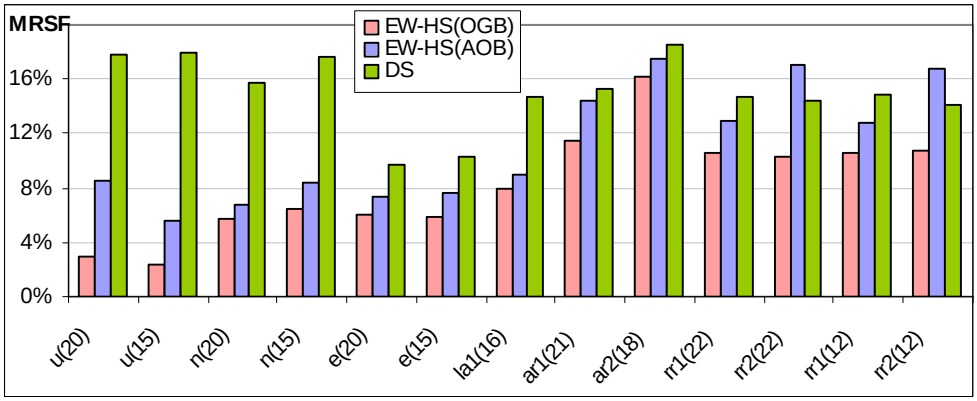


Abbildung 5.23: MRSF des EW-HS bei OGB und AOB-NS sowie DS jeweils bei 1%-Anfragen.

Zuletzt wurde der Equi-Width Histogrammselektivitätsschätzer (EW-HS) mit dem Average-Shifted Histogrammselektivitätsschätzer (ASHS) verglichen. Durch Variation der Anzahl Shifts wurde herausgefunden, daß eine Anzahl s von 10 Shifts i.a. genügend ist für alle Testdatensätze. Die Ergebnisse wurden verglichen, indem als Anzahl von Bins das bei Anwendung der Normalskalierungsregel für Histogramme ermittelte k gewählt wurde. In nahezu allen Fällen

schneidet der ASHS besser ab als der EW-HS, vgl. Abbildung 5.24. Daher ist im univariaten Fall, wo der zusätzliche Rechen- und Speicheraufwand noch vertretbar ist, der ASHS den untersuchten Histogrammselektivitätsschätzern (EW-, ED-, MD-HS) vorzuziehen.

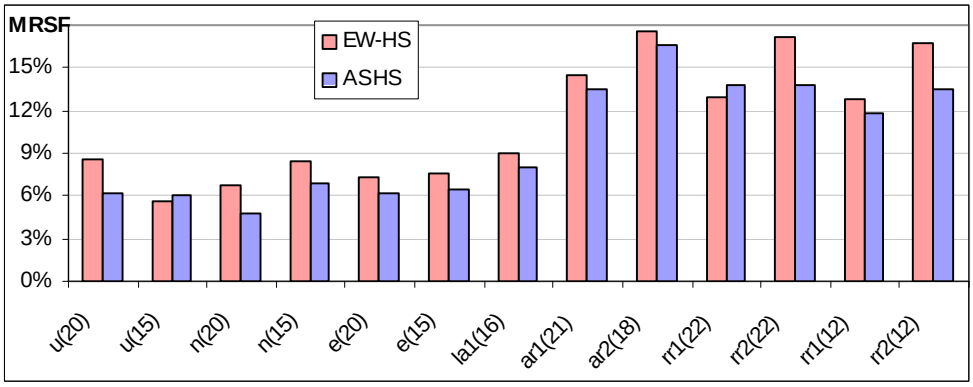


Abbildung 5.24: MRSF des EW-HS und des ASHS bei 1%-Anfragen.

Zusammenfassend lassen sich die Ergebnisse dieses Abschnittes wie folgt festhalten: Von den drei verschiedenen Histogrammselektivitätsschätzern (Equi-Width, Equi-Depth, Max-Diff) liefert der EW-HS i.a. die besseren Ergebnisse bei Vorliegen einer optimalen gefundenen Anzahl von Bins. Das EW-HS hat des weiteren gegenüber den anderen HS den Vorteil, daß ein Verfahren zur Bestimmung der asymptotisch optimalen Binweite vorliegt. Dieses Verfahren liefert durchaus akzeptable Ergebnisse - die von einer Ausnahme abgesehen - bei weitem besser als bei der direkten Selektivitätsschätzung sind. Bessere Verfahren zur Schätzung der AOB werden im Zusammenhang der Kernselektivitätsschätzer im Abschnitt 5.5.5 diskutiert. Besser noch als die soeben besprochenen HS schneidet der Average-Shifted Histogrammselektivitätsschätzer ab. Obwohl bei diesem Verfahren die Schätzung der AOB sowie als weiterem Parameter die Schätzung der optimalen Anzahl von Shifts gewisse Probleme bereiten, ist dieses Verfahren den übrigen diskutierten Histogrammselektivitätsschätzern im univariaten Fall vorzuziehen.

5.5.4 Ergebnisse der verschiedenen Kernselektivitätsschätzer

In diesem Abschnitt werden Selektivitätsschätzer auf Grundlage von Kernfunktionen experimentell untersucht. Dazu wurde der univariate Kernselektivitätsschätzer (KS) ohne Randbehandlung (KS-O) sowie mit der Spiegelung (KS-S) als auch mit speziellen Randkernen (KS-R) als Randbehandlung implementiert und mit den Testdaten getestet. Als Kernfunktion wurde immer der Epanechnikow-Kern bzw. in der Variante als Randkern der Epanechnikow-Randkern von [Dong & Simonoff 94] verwendet. Das Problem der Bandbreitenbestimmung wird im nächsten Abschnitt separat ausführlich untersucht.

Zuerst wird das in Kapitel 3.4.2 adressierte Randproblem für Kernschätzer untersucht. Dazu werden in Abbildung 5.25 der relative Selektivitätsfehler für KS-O, KS-S und KS-R bei 1%-Anfragen verglichen. Die Bandbreite wurde dabei variabel gelassen, so daß die Abbildung die

Ergebnisse bei jeweils optimaler gefundener Bandbreite OGB repräsentiert. Dabei zeigt sich, daß Randprobleme durch Kernselektivitätsschätzer mit geeigneter Randbehandlung zufriedenstellend gelöst werden können. Bei den künstlichen Datensätzen mit Rand (u(p) und e(p)) ergibt sich eine deutliche Verbesserung der Schätzung. Bei den gleichverteilten Daten u(20) verbessert sich der relative Selektivitätsfehler von 8,3% ohne Randbehandlung (gefundene optimale Bandbreite h = 120.000) auf 2,9% (Bandbreite $h = h_{max}$ =400.000) bei Anwendung der Spiegelung. Lediglich bei den normalverteilten Daten n(15) ergab sich eine leichtere Verschlechterung bei der Spiegelung, da hier bei großer Bandbreite zu viel Masse an den äußeren Rand gespiegelt wird. Bei den realen Datensätzen ergab sich nur bei den la(16) Daten eine leichte Verbesserung, bei den anderen realen Datensätzen blieben die Ergebnisse gleich. Es ist daher zu vermuten, daß bei diesen keine Daten am Rand vorliegen, so daß eine explizite Randbehandlung nicht nötig wäre. Vergleicht man die Verfahren mit Spiegelung und mit speziellem Randkern miteinander, so ergeben sich überraschenderweise kaum Unterschiede.

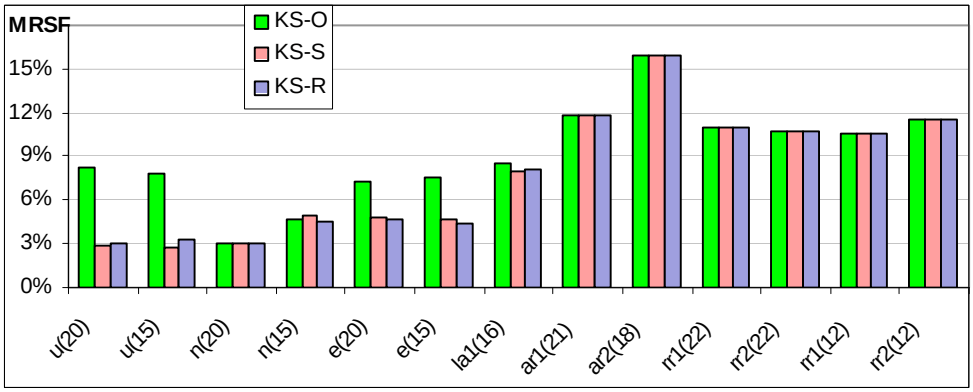


Abbildung 5.25: MRSF für KS bei 1%-Anfragen mit OGB.

Abbildung 5.26 zeigt deutlich die Reduktion des MRSF am Rand durch geeignete Randbehandlung. Die Kurven zeigen den MRSF der unterschiedlichen Kernselektivitätsschätzer (ohne Randbehandlung KS-O, mit Spiegelung KS-S, mit speziellen Randkernen KS-R) bei 1%-Anfragen auf dem gleichverteilten Datensatz u(20) in Abhängigkeit vom Mittelpunkt der Anfragebereiche. Im mittleren Bereich [h, $r - h$], $r = 2^{20} - 1$ ist der MRSF bei allen 3 Kurven gleich. Zum Rand hin steigt der MRSF beim KS-O jedoch stark an, während er beim KS-S und KS-R im Randbereich auf etwa dem gleichen Niveau wie im mittleren Bereich bleibt.

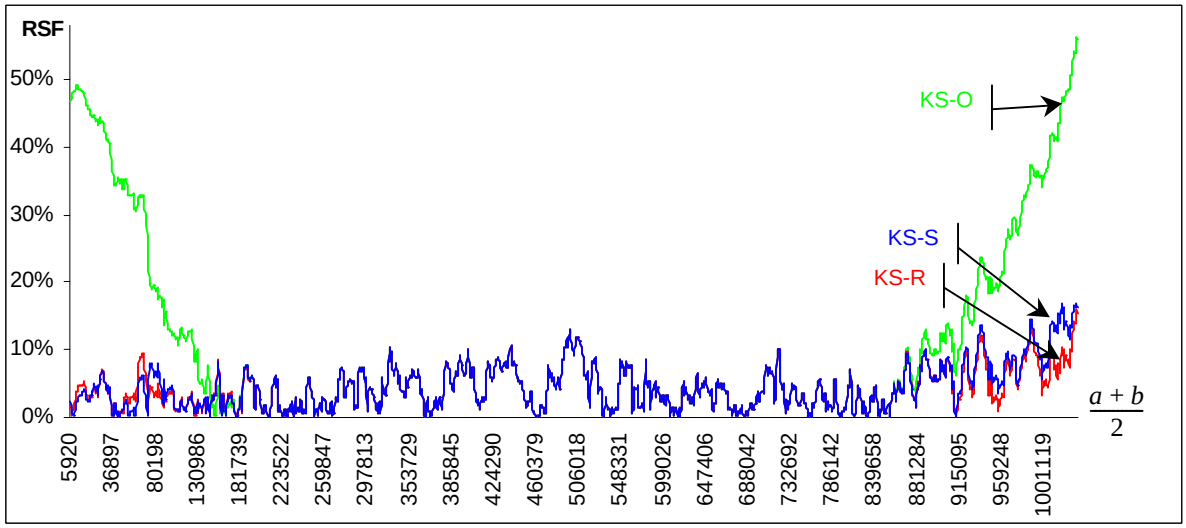


Abbildung 5.26: RSF verschiedener KS abhängig von der Position der Mittelpunkte der Anfragen bei u(20)-Testdaten und 1%-Anfragen.

Abbildung 5.27 zeigt die Auswirkungen der Randbehandlung auf die Güte der Schätzung bei gleichverteilten Daten (u(20)) und variabler Bandbreite *h*. Erfolgt die Kernselektivitätsschätzung ohne Randbehandlung (KS-O), so wird der Randfehler ab einer gewissen Bandbreite so groß, daß der relative Selektivitätsfehler wieder ansteigt, während er bei der Kernselektivitätsschätzung mit Randbehandlung bei wachsender Bandbreite monoton fällt. Dies entspricht den Erwartungen, da die optimale Bandbreite bei gleichverteilten Daten unendlich ist.

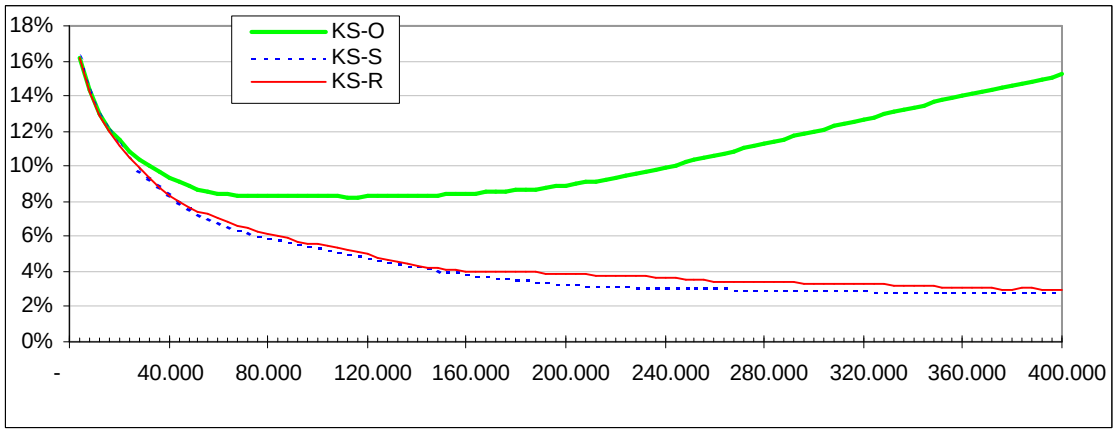


Abbildung 5.27: RSF der Kernselektivitätsschätzung mit und ohne Randbehandlung bei u(20)-Daten abhängig von der Bandbreite *h*.

Da sich die Ergebnisse durch die Verwendung von Kernselektivitätsschätzer mit Randbehandlung kaum verschlechtern aber durchaus erheblich verbessern können, werden im folgenden hauptsächlich Kernselektivitätsschätzer mit Randbehandlung betrachtet.

Abbildung 5.28 vergleicht nun den Kernselektivitätsschätzer mit Randkern mit der direkten Selektivitätsschätzung. Hierbei wurde bei allen Testdatensätzen jeweils die optimale gefundene Bandbreite herangezogen. Es zeigt sich deutlich die Überlegenheit der Kernselektivitätsschätzer insbesondere bei den künstlichen Datensätzen.

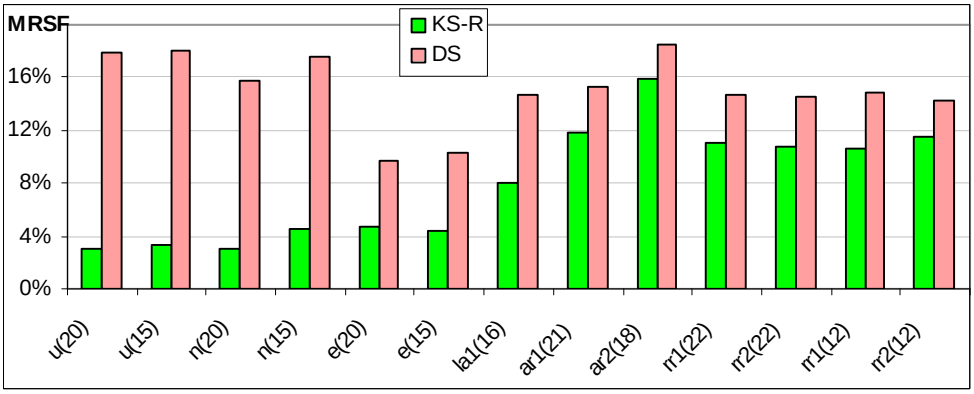


Abbildung 5.28: MRSF für KS-R(OGB) sowie für DS bei 1%-Anfragen.

Als nächstes werden die Ergebnisse der Kernselektivitätsschätzung (mit Randbehandlung) mit denen der Histogrammselektivitätsschätzer aus vorigem Abschnitt verglichen. Hier zeigt sich überraschenderweise, daß die Ergebnisse bei künstlichen und realen Datensätze sehr unterschiedlich sind. Abbildung 5.29 zeigt daher zunächst nur die Ergebnisse für die künstlichen Testdaten. Im linken Teil a) sind die Ergebnisse bei optimaler gefundener Bandbreite (OGB) präsentiert, während bei den Ergebnisse in der rechten Grafik b) die mittels Normalskalierungsregel geschätzte Bandbreite verwendet wurde. Abgesehen von den gleichverteilten Datensätzen zeigt sich in allen Fällen die Überlegenheit der Kernselektivitätsschätzung gegenüber der Histogrammselektivitätsschätzung. Dies gilt insbesondere, wenn die (asymptotisch) optimale Bandbreite nicht bekannt ist, sondern durch die Normalskalierungsregel geschätzt werden muß (AOB-NS). Hierbei schneiden auch die gleichverteilten Testdaten bei der Kernselektivitätsschätzung besser ab als bei der Histogrammselektivitätsschätzung (u(20): KS-R 4,1% und EW-HS 8,5%).

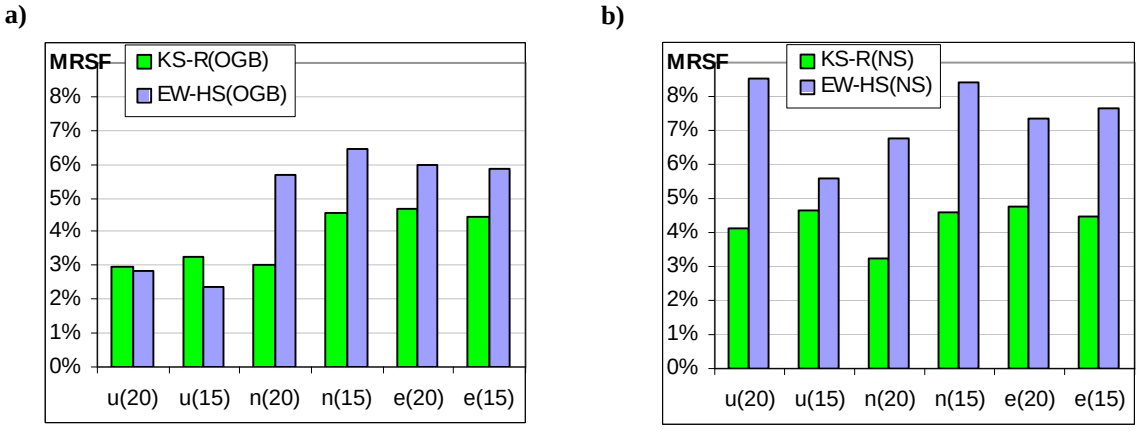


Abbildung 5.29: MRSF von KS-R und EW-HS bei künstlichen Testdaten: **a)** mit OGB, **b)** mit AOB-NS.

Ein anderes Bild zeigt sich allerdings bei den realen Testdaten, wie in Abbildung 5.30 dargestellt. Hier zeigt sich bei der optimalen gefundenen Bandbreite, daß der Histogrammselektivitätsschätzer leicht besser als der Kernselektivitätsschätzer abschneidet. Bei der mittels der Normalskalierungsregel bestimmten Bandbreite sind die Ergebnisse des Histogrammselektivitätsschätzers teilweise sogar deutlich besser als beim Kernselektivitätsschätzer (ar1(21): MRSF(KS-R) = 22,7% und MRSF(EW-HS) = 14,4%). Der Grund für dieses unterschiedliche Verhalten bei künstlichen und realen Testdaten ist vermutlich darin zu suchen, daß bei den realen Datensätzen Sprungstellen in den Testdaten vorliegen, die der Annahme der stetigen Dichtefunktion widersprechen. Diese Vermutung wird durch die Ergebnisse des Hybridschätzers bestätigt, bei dem speziell auf die Sprungstellenproblematik eingegangen wird (siehe hierzu Abschnitt 5.5.6).

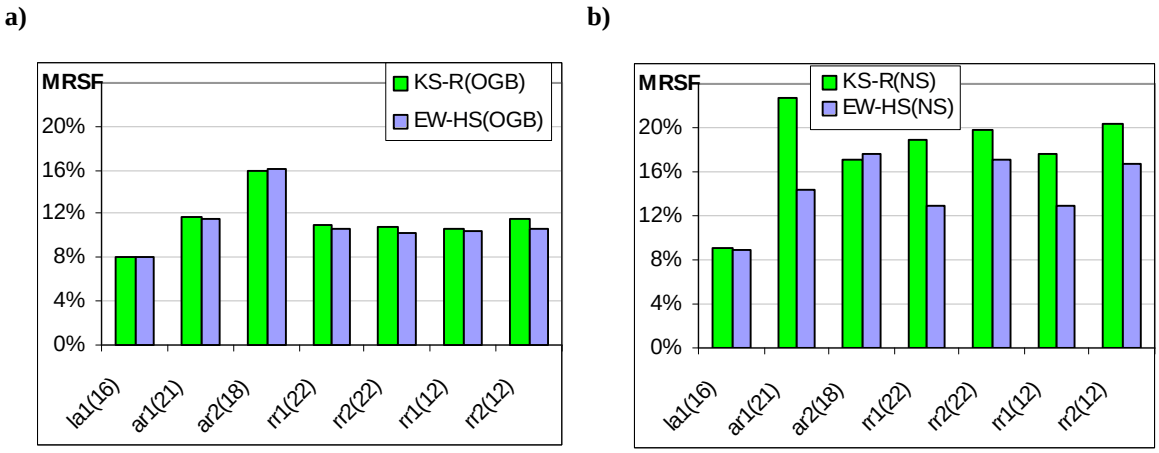


Abbildung 5.30: MRSF von KS-R und EW-HS bei realen Testdaten: **a)** mit OGB, **b)** mit AOB-NS.

Zuletzt sei der Kernselektivitätsschätzer mit Randkern (KS-R) mit dem Average-Shifted Histogrammselektiviätsschätzer (ASHS) verglichen. Dabei werden die Ergebnisse des KS-R sowohl bei optimaler gefundener Bandbreite (OGB) als auch bei mittels Normalskalierungsregel geschätzter asymptotisch optimaler Bandbreite (AOB-NS) dargestellt. Die Bandbreite des ASHS ist analog zum Vergleich mit den Histogrammselektivitätsschätzern gewählt. Wegen des unterschiedlichen Ergebnisses der Kernselektivitätsschätzung bei künstlichen und realen Testdaten, werden die Ergebnisse wieder in Abbildung 5.31 in zwei getrennten Grafiken präsentiert. Hierbei zeigt sich ein ähnliches Bild wie schon beim Vergleich der KS-R mit der Histogrammselektivitätsschätzung. Bei den künstlichen Testdaten hat die KS-R deutlich bessere Ergebnisse als die ASHS. Dies trifft auch für die KS-R bei geschätzter AOB zu. Bei den realen Testdaten verliert die KS-R im Vergleich zur ASHS wieder deutlich an Qualität. Nur bei optimaler gefundener Bandbreite sind die Ergebnisse der KS-R i.a. besser als bei der ASHS - bei den la1(16)-Daten sind sogar diese leicht schlechter (KS-R(OGB) 8,05%, ASHS 8,03%).

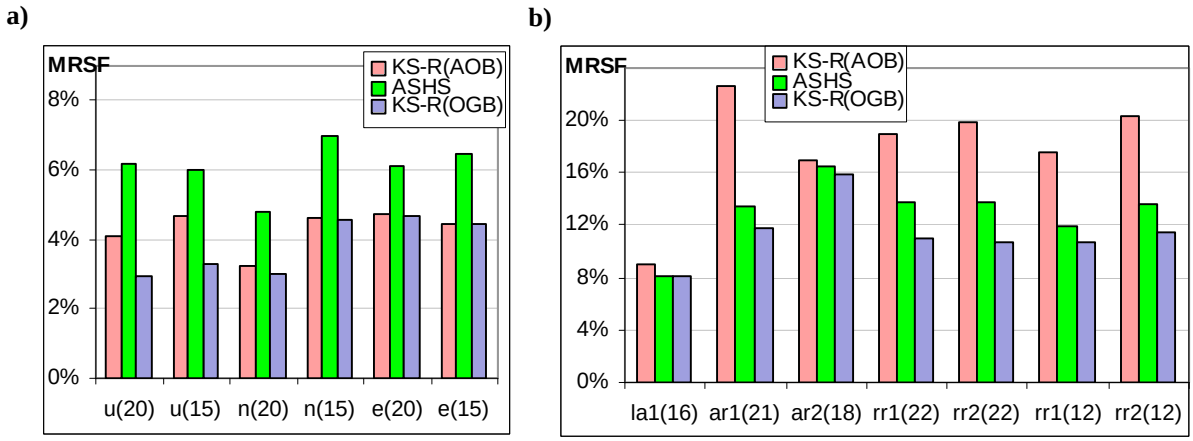


Abbildung 5.31: MRSF von KS-R bei OGB und AOB-NA und von ASHS.

Die Ergebnisse dieses Abschnitts zeigen, daß der Kernselektivitätsschätzer eine geeignete Methode zur Selektivitätsschätzung darstellt. Da es bei manchen Datenverteilungen zu Randproblemen kommen kann, ist in diesen Fällen eine Randbehandlung sehr wichtig. Die Ergebnisse mit den Testdaten zeigen weiterhin, daß die Randbehandlung in solchen Fällen zu wesentlich besseren Resultaten führt, und es i.a. bei Datenverteilungen ohne Masse an den Rändern höchsten zu einer leichten Verschlechterung der Ergebnisse kommt (normalverteilte Testdaten). Daher ist in praktischen Anwendungen ohne vorherige Kenntnis der Verteilung an den Rändern immer die Kernselektivitätsschätzung mit Randbehandlung zu verwenden.

Beim Vergleich der Kernselektivitätsschätzung mit der Histogrammselektivitätsschätzung und der Average Shifted Histogrammselektivitätsschätzung zeigen sich deutlich unterschiedliche Ergebnisse der Kernselektivitätsschätzung bei künstlichen und bei realen Testdaten. Während sich bei künstlichen Testdaten gute und weitaus bessere Ergebnisse als bei der Histogrammselektivitätsschätzung und auch der Average Shifted Histogrammselektivitätsschätzung ergeben, liegen bei realen Testdaten wesentlich schlechtere Ergebnisse vor. Um trotzdem Kernschätzer auch bei realen Datensätzen sinnvoll einsetzen zu können, ist somit eine besondere Behandlung bei realen Datensätzen notwendig. Ein solcher Ansatz wird mit dem Hybridselektivitätsschätzer verfolgt, deren Ergebnisse in Abschnitt 5.5.6 präsentiert werden. Die dortigen Ergebnisse bestätigen die zuvor gemachten Überlegungen.

Im folgenden werden zunächst die Ergebnisse verschiedener Verfahren zur Schätzung der Bandbreite bei Kernschätzern diskutiert.

5.5.5 Bewertung der verschiedenen Verfahren zur Bandbreitenbestimmung

In diesem Abschnitt werden verschiedene der in Kapitel 4.4 vorgestellten Verfahren zur Schätzung der (asymptotisch optimalen) Bandbreite (AOB) bei Kernschätzern anhand der Ergebnisse mit den Testdaten diskutiert. Untersucht wurde zum einen die Normalskalierungsregel (-NS) wie in 4.4.1 vorgestellt, da sie zu den bekanntesten und einfachsten Verfahren zur Schätzung der AOB gehört. Als weiteres modernes Verfahren wurde die direkte Plug-In Methode 2. Stufe

(-DPI2) wie in 4.4.3 vorgestellt angewendet. Des weiteren wurde als ein adaptives das in 4.4.4 vorgestellte Verfahren untersucht (-AD). Die Ergebnisse der optimalen gefundenen Bandbreite (OGB) dienen dabei als Referenzwerte.

Tabelle 5.9 listet zum Vergleich für alle Testdaten die durch die verschiedenen Verfahren bestimmten Bandbreiten auf. Dabei wurde die OGB beim Kernselektivitätsschätzer mit Randbehandlung ermittelt. Da die Bandbreite beim adaptiven Verfahren variabel ist, findet sie sich nicht in der Tabelle wieder. Die Tabelle zeigt, daß die AOB-NS sowohl bei den exponentialverteilten als auch bei den realen Testdaten viel zu hoch ist, so daß eine deutliche Überglättung zu erwarten ist. Dieses Verhalten findet sich in deutlich abgeschwächter Form bei der direkten Plug-In Regel wieder. Bei den gleichverteilten Testdaten wäre eine möglichst hohe Binweite optimal, so daß hier die Normalskalierungsregel ausnahmsweise besser als die direkte Plug-In Regel ist. Wirklich nahe an die OGB kommen die beiden Verfahren nur bei den normalverteilten Testdaten, wobei die AOB-DPI2 etwas näher an der OGB liegt. Von daher scheint i.a. die direkte Plug-In-Regel das bessere Schätzverfahren gegenüber der Normalskalierungsregel darzustellen. Dies ist durch Vergleich der relativen Selektivitätsfehler im folgenden zu überprüfen.

Testdatensatz	**OGB**	**AOB-NS**	**AOB-DPI2**
u(20)	> 400.000	155.527	91.573
u(15)	> 10.000	4.853	3.193
n(20)	136.800	118.524	124.446
n(15)	3.840	3.690	3.734
e(20)	32.000	36.715	15.490
e(15)	1.040	1.147	484
la(16)	2.304	9.626	4.962
ar1(21)	20.400	99.833	47.480
ar2(18)	2.940	36.785	22.232
rr1(22)	84.000	473.164	248.454
rr2(22)	60.000	443.597	243.296
rr1(12)	84	462	243
rr2(12)	56	424	236

Tabelle 5.9: OGB und geschätzte AOB für KS.

Aufgrund der übersichtlicheren Darstellung aber auch aufgrund des im vorigen Abschnitt beobachteten unterschiedlichen Verhaltens bei künstlichen und realen Testdaten werden diese im folgenden getrennt betrachtet.

Da beim adaptiven Verfahren keine spezielle Randbehandlung existiert, seien zunächst die Ergebnisse der Kernschätzung mit verschiedenen Bandbreitenschätzern ohne Randbehandlung betrachtet. Abbildung 5.32 zeigt die Ergebnisse bei den künstlichen Testdaten. Die beiden Verfahren mit fester Bandbreite zeigen bei den normalverteilten Daten wie erwartet vergleichbare

Ergebnisse, auch die Bandbreite unterscheidet sich nur unwesentlich. Bei den anderen künstlichen Testdaten liefert das direkte Plug-In Verfahren eine bessere Bandbreite mit geringerem MRSF. Dies trifft insbesondere bei den exponential-verteilten Testdaten zu. Dies sind auch die einzigen künstlichen Testdaten, bei denen das adaptive Bandbreitenverfahren zu einem geringerem MRSF führt als die anderen Verfahren. Der MRSF ist sogar geringer als beim KS-O mit optimaler gefundener fester Bandbreite. Dies ist dadurch erklärbar, daß durch die hohe Schiefe im linken Bereich mit hoher stark fallender Dichte eine höhere Bandbreite erforderlich ist als im flachen rechten Teil.

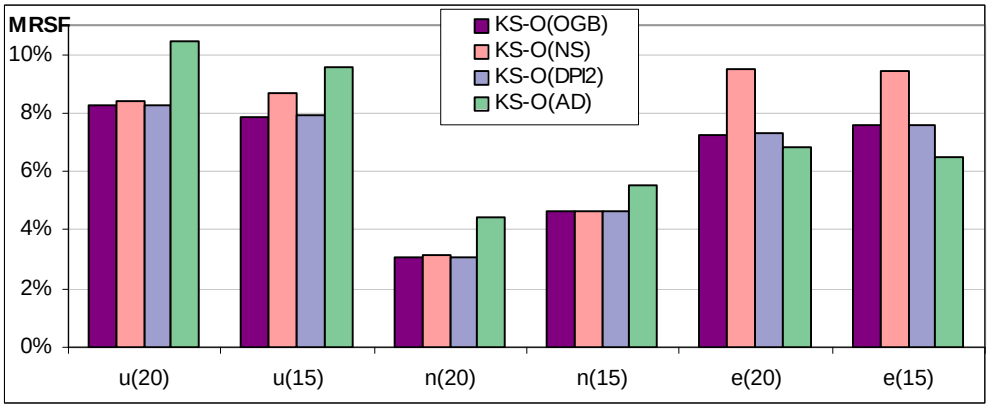


Abbildung 5.32: MRSF für KS-O bei unterschiedlichen Bandbreitenverfahren und künstlichen Testdatensätzen.

Die nächste Abbildung 5.33 zeigt die Ergebnisse bei Kernselektivitätsschätzung mit Randbehandlung mit verschiedenen Bandbreitenschätzern inkl. adaptiver Bandbreitenwahl (ohne Randbehandlung) und künstlichen Testdaten. Bei den normalverteilten Testdaten gibt es nur einen geringen Unterschied zu der vorigen Abbildung, da diese Daten so gut wie keine Dichte am Rand aufweisen. Bei den gleichverteilten Testdaten zeigt sich eine Verbesserung insbesondere bei der Normalskalierungsregel. Letztere erzeugt i.a. die größere Bandbreite, was sich in diesem Fall positiv auswirkt, da dies der optimalen Bandbreite bei der Gleichverteilung näher kommt. Auch bei den exponentiell verteilten Testdaten ergibt sich eine deutliche Verbesserung unter Verwendung der Randbehandlung. Bei nahezu allen künstlichen Testdaten liefert das Verfahren mit der Normalskalierungsregel die besten Ergebnisse. Das adaptive Verfahren ohne Randbehandlung schneidet hier am schlechtesten ab.

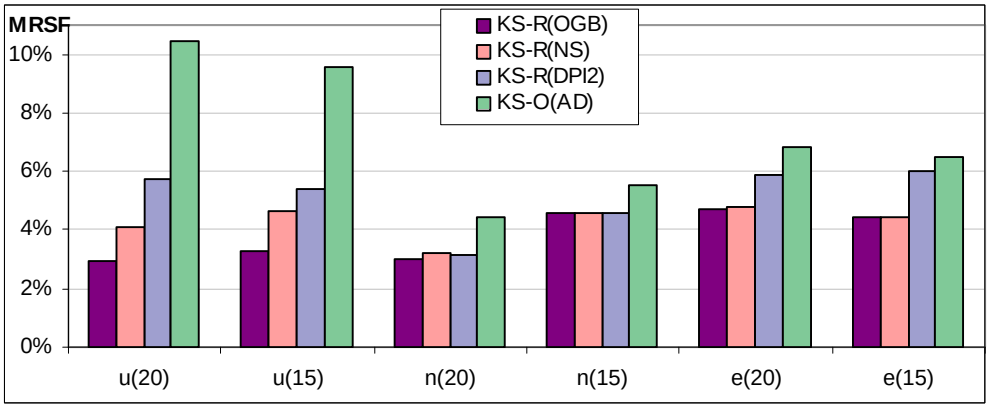


Abbildung 5.33: MRSF für KS-R bei unterschiedlichen Bandbreitenverfahren bzw. KS-O(AD) und künstlichen Testdatensätzen.

Abbildung 5.34 zeigt nun die Ergebnisse bei den realen Testdaten ohne Randbehandlung. Bei den realen Datensätzen zeigt sich ein von den künstlichen Datensätzen verschiedenes aber einheitliches Bild. In allen Fällen liefert die direkte Plug-In Regel 2. Stufe (DPI2) teilweise deutlich bessere Ergebnisse als die Normalskalierungsregel. Das adaptive Verfahren liefert vergleichbare Ergebnisse wie die DPI2-Regel. Bei einigen Testdaten schneidet das adaptive Verfahren besser ab (ar1(21), rr2(22), rr2(12)), bei anderen die DPI2-Regel (la1(16), ar2(18),rr1(22)). Beide Verfahren sind daher der Normalskalierungsregel bei realen Testdaten in jedem Fall vorzuziehen. Eine Bevorzugung des einen oder anderen Verfahrens ist anhand der vorliegenden Ergebnisse nicht entscheidbar.

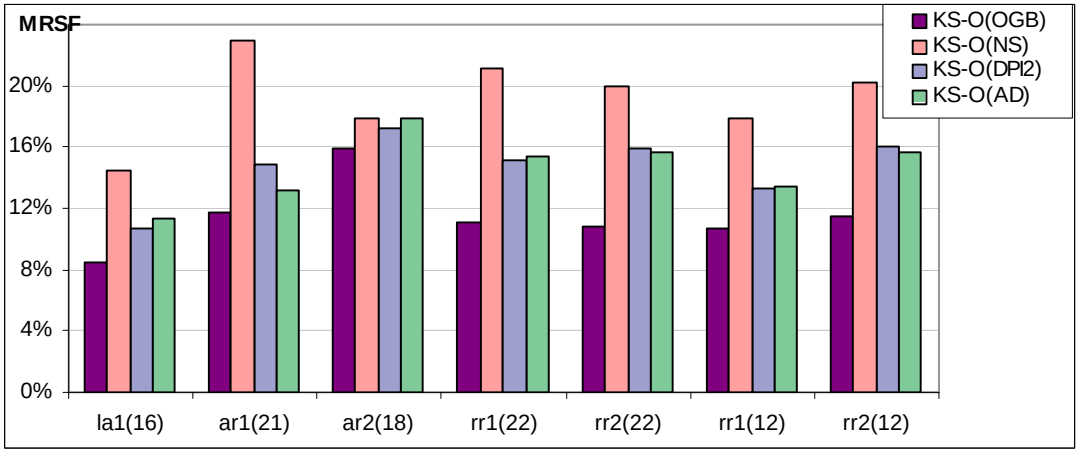


Abbildung 5.34: MRSF für KS-R bei unterschiedlichen Bandbreitenverfahren und realen Testdatensätzen.

In Abbildung 5.35 sind letztendlich die Ergebnisse bei Kernschätzung mit Randbehandlung mit verschiedenen Bandbreitenschätzern inkl. adaptiver Bandbreitenwahl (ohne Randbehandlung) und künstlichen Testdaten dargestellt. Die Ergebnisse sind ähnlich zur vorherigen Abbildung,

da die untersuchten Testdaten kaum Dichte am Rand aufweisen, so daß die Randbehandlung nur zu einer minimalen Verbesserung bei einigen Testdaten führt.

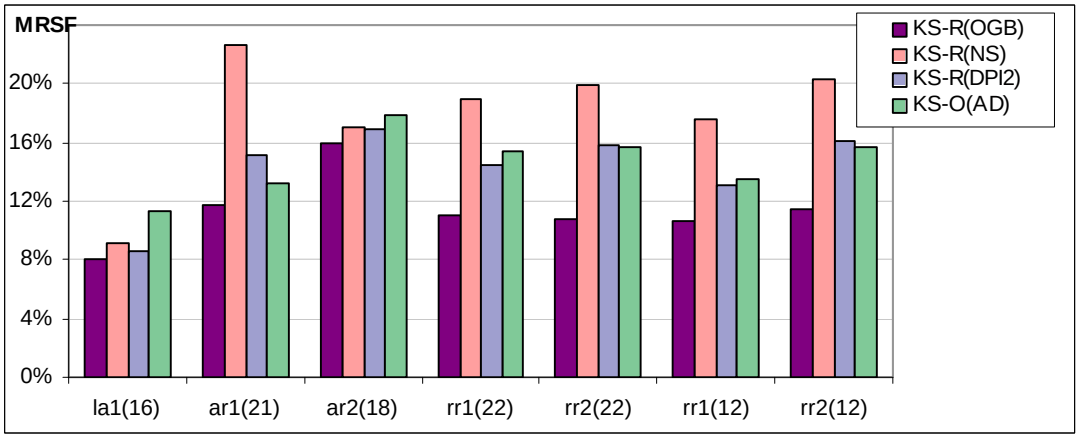


Abbildung 5.35: MRSF für KS-R bei unterschiedlichen Bandbreitenverfahren und realen Testdatensätzen.

Die Experimente mit dem Kernselektivitätsschätzer bestätigen, daß die Wahl der Bandbreite einen entscheidenden Einfluß auf die Güte der Schätzung hat. Da in einem DBMS die Wahl der Bandbreite automatisch erfolgen muß, sind entsprechend automatische Verfahren zur Bandbreitenbestimmung notwendig. Dabei handelt es sich um kein triviales Problem, und die verschiedenen Verfahren zeigen unterschiedliche Ergebnisse. Bei künstlichen Datensätzen liefern die Verfahren i.a. brauchbare bis sehr gute Ergebnisse. Bei realen Datensätzen können die Ergebnisse sehr viel schlechter ausfallen, wobei allerdings der Kernselektivitätsschätzer selbst bei OGB nicht viel besser ist. Diesem allgemeinen Problem wird versucht durch den Hybridschätzer beizukommen, die Ergebnisse sind im nächsten Abschnitt wiedergegeben.

Vergleicht man die verschiedenen Schätzverfahren der AOB miteinander, so bringt die einfachere Normalskalierungsregel bei den künstlichen Testdaten brauchbare Ergebnisse. Dies sieht jedoch bei den realen Testdaten völlig anders aus. Hier stechen sowohl die DPI2-Regel als auch das adaptive Verfahren die Normalskalierungsregel aus. Eine Bevorzugung des einen oder anderen Verfahrens ist anhand der Ergebnisse nicht entscheidbar.

5.5.6 Bewertung der Ergebnisse des Hybridselektivitätsschätzers

Die in Abschnitt 5.5.4 durchgeführten Experimente mit Kernselektivitätsschätzern zeigten schlechte Ergebnisse bei realen Testdaten. Dabei ist bei der Betrachtung der Einzelfehler der Anfragen das Auftreten von Extremstellen zu beobachten. Vgl. dazu Abbildung 5.36, in der die Differenz von wahrer und geschätzter Selektivität an den Mittelpunkten der Anfragenintervalle bei den ar2(18)-Testdaten aufgetragen wurde. Im folgenden wird angenommen, daß diese Extrema durch Sprungstellen in der den Daten zugrunde liegenden Dichte begründet sind. In

diesem Abschnitt wird der eigens hierfür entwickelte Hybridselektivitätsschätzer aus Kapitel 3.5 auf solche Daten angewendet.

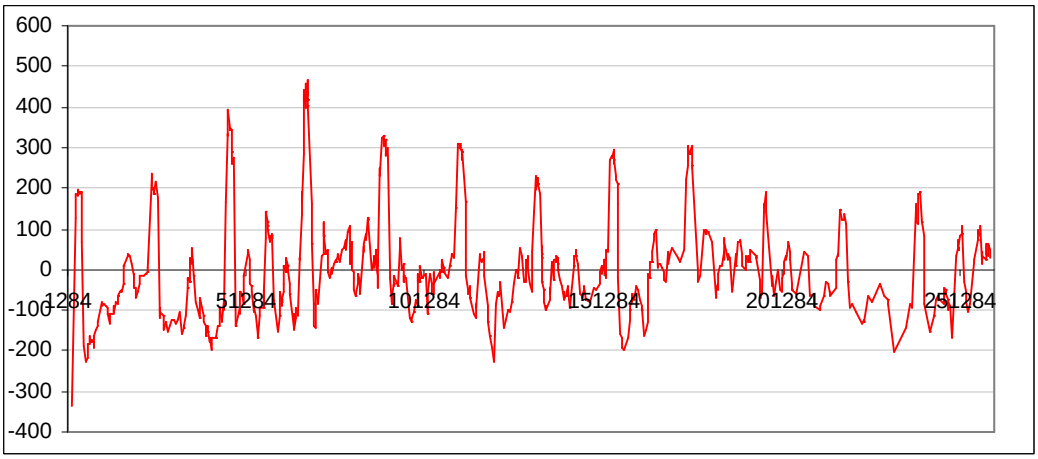


Abbildung 5.36: Fehler des KS-R bei ar2(18)-Testdaten und AOB-NS mit h = 36.785.

Dazu wird der Hybridselektivitätsschätzer zunächst mit den Testdaten xu(15) und xe(15) untersucht, bei denen es sich um künstliche Testdaten mit (insgesamt 8 bekannten) künstlichen Sprungstellen handelt, vgl. Kapitel 5.3.4. Abbildung 5.37 zeigt die Einzelfehler des Kernselektivitätsschätzers mit Randkern und des Hybridselektivitätsschätzers bei den xe(15)-Testdaten. Dort ist wiederum die Differenz aus wahrer und geschätzter Selektivität an den Mittelpunkten der Anfrageintervalle aufgetragen. Die Bandbreite des Kernselektivitätsschätzer wurde mit der Normalskalierungsregel geschätzt und beträgt h = 3.746. Für den Hybridselektivitätsschätzer wurde ein Histogramm mit 9 Bins erzeugt, deren Grenzen den (bekannten) Sprungstellen (s. Tab. 5.6) entsprechen. Die Abbildung zeigt 9 Extrema in der Fehlerkurve beim Kernselektivitätsschätzer (rote normale Linie), deren Lage in etwa der Lage der Sprungstellen entspricht. Dagegen ergibt sich eine deutliche Reduktion des RSF beim Hybridselektivitätsschätzer bereits ohne Randbehandlung (gestrichelte grüne Linie). Dieser Fehler wird weiter reduziert durch die Benutzung von Randkernen an den Rändern der Histogrammbins.

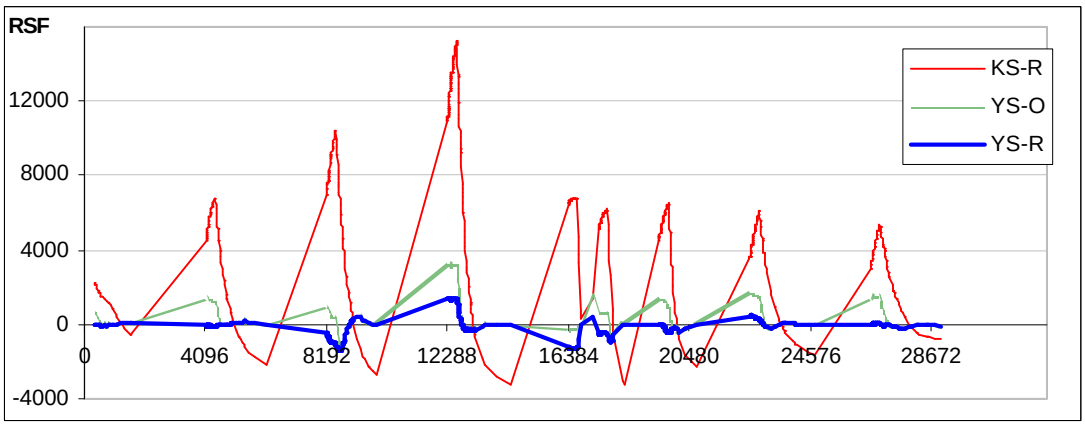


Abbildung 5.37: Fehler der KS-R bei xe(15)-Daten und AOB

Abbildung 5.38 zeigt den mittleren relativen Selektivitätsfehler sowohl für die xu(15)- als auch die xe(15)-Testdaten bei verschiedenen Selektivitätsschätzern. Dabei wurde bei allen Schätzern die Bandbreite mit Normalskalierungsregel geschätzt. Tabelle 5.10 zeigt die jeweiligen Bandbreiten für den EW-HS und den KS-R. Die Histogramme des YS-R wurden mit den exakten Sprungstellen gesetzt. Die xu(15)-Testdaten zeigen bei dem YS nur leichte Verbesserungen (MRSF(YS-R) = 8,7%) gegenüber dem EW-HS (MRSF = 9,0%), was darin begründet ist, daß HS bei gleichverteilten Daten aufgrund der Gleichverteilungsannahme innerhalb eines Intervalls immer besser sind. Durch die hohe Anzahl von Bins gegenüber den Sprungstellen kommt es kaum zu Sprüngen innerhalb der Histogrammbins. Bei den xu(15) Testdaten wurde daher noch ein HS mit dem exakten Histogramm angewendet. Der MRSF reduzierte sich dabei sogar deutlich auf 3,1%. I.a. ist jedoch bei realen Daten davon auszugehen, daß die Daten nicht ("stückweise") gleichverteilt sind. Eine deutliche Verbesserung des MRSF ergibt sich daher bei den xe(15)-Testdaten. Hier reduziert sich der MRSF von über 62% beim EW-HS auf 6,8% beim YS-R.

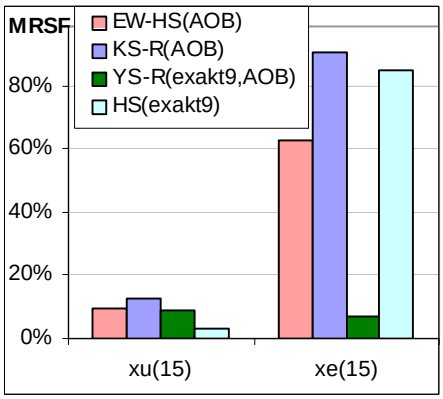


Abbildung 5.38: MRSF für den EW-HS, den KS-R und den YS bei den xu(15)- und den xe(15)-Testdaten.

	Anzahl Klassen beim EW-HS	**Bandbreite des KS-R**
xu(15)	21	3.822
xe(15)	19	3.746

Tabelle 5.10: AOB-NS bei EW-HS bzw. KS-R für xu(15)- und xe(15)-Testdaten

Die zuvor angestellten Beobachtungen und (heuristischen) Überlegungen sollen nun auch reale Testdaten angewendet werden. Dazu wurden die Sprungstellen anhand der Extrema der Fehlerkurve gesetzt, vgl. Abb. 5.39 bei den ar2(18)-Testdaten. Es handelt sich um die gleiche Kurve wie in Abb. 5.37 mit dem Unterschied, daß hier die zur Bildung der Bingrenzen herangezogenen Maxima durch kleine Quadrate markiert wurden. Im angegebenen Beispiel ergibt sich für die ar2(18)-Testdaten ein Histogramm mit 12 Bins. Abb. 5.40 zeigt das gleiche Vorgehen bei den la1(16)-Testdaten. Hier ergibt sich ein Histogramm mit lediglich 2 Bins. Bei den rr2(22)-Testdaten wurden die Bingrenzen anhand der Maxima so gewählt, daß sich ein Histogramm mit

8 Bins ergibt, vgl. Abb. 5.41. Tabelle 5.11 zeigt die Anzahl der Histogrammbins für die jeweiligen realen Testdaten, die anschließend für den YS verwendet wurden.

Testdaten	**la1(16)**	**ar1(21)**	**ar2(18)**	**rr1(22)**	**rr2(22)**
Anzahl Bins	2	11	12	11	8

Tabelle 5.11: Anzahl Bins der für die Hybridselektivitätsschätzung verwendeten Histogramme.

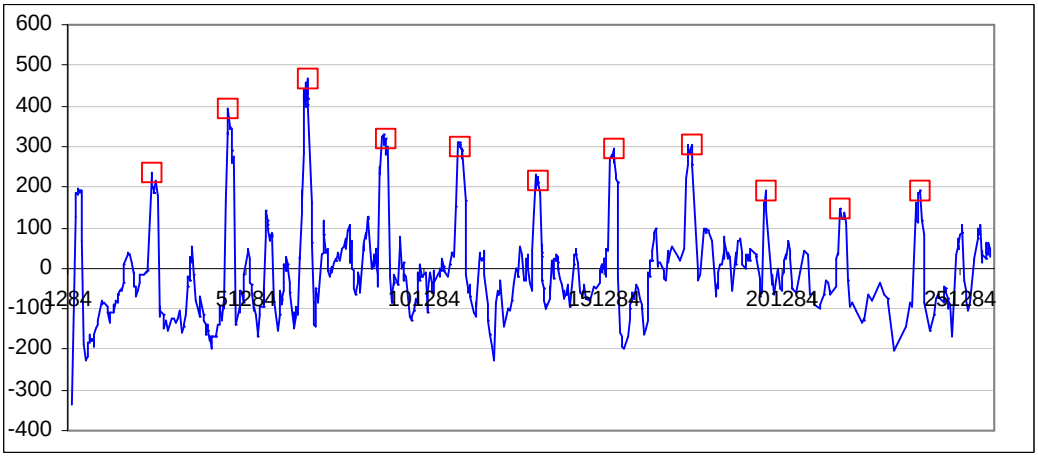


Abbildung 5.39: Fehler der KS-R bei ar2(18)-Daten und AOB.

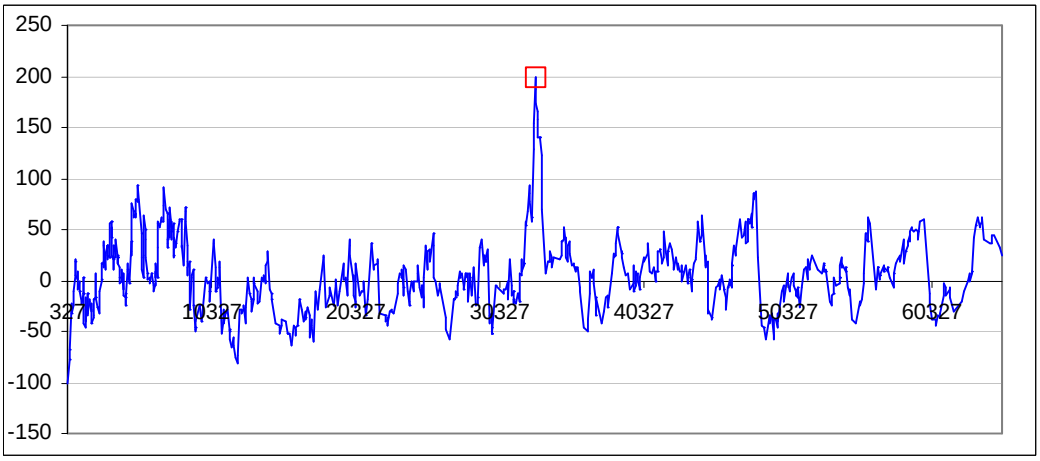


Abbildung 5.40: Fehler der KS-R bei la(16)-Daten und AOB.

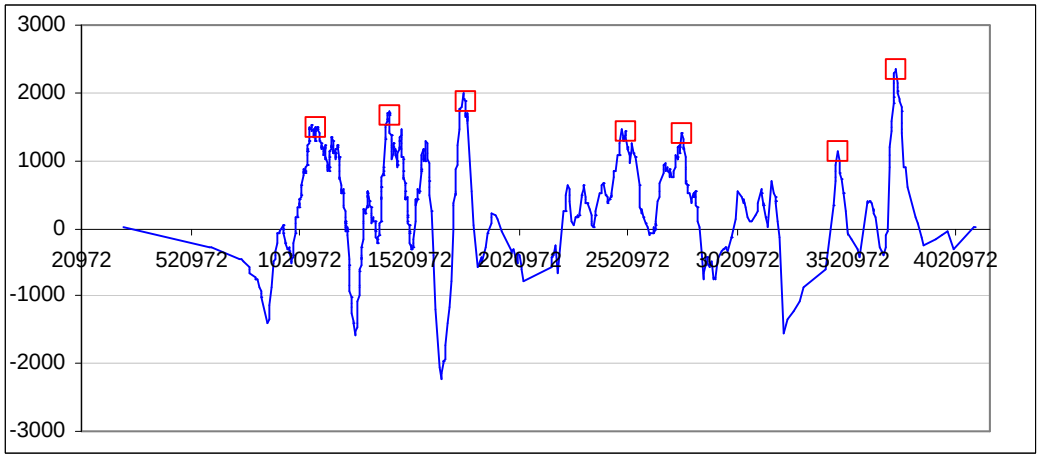


Abbildung 5.41: Fehler der KS-R bei rr2(22)-Daten und AOB.

Abbildung 5.42 zeigt die Ergebnisse des Hybridselektivitätsschätzer YS-R(x) im Vergleich zu den beiden verwandten Selektivitätsschätzern, dem Histogrammselektivitätsschätzer mit gleichem zugrundegelegten Histogramm HS(x) sowie dem Kernselektivitätsschätzer mit Randbehandlung KS-R, bei den realen Testdaten. Sowohl beim KS-R als auch beim YS-R wurde bei der Kernschätzung die Normalskalierungsregel zur Schätzung der Bandbreite verwendet. Zudem wurden die Ergebnisse mit dem KS-R mit optimaler gefundener Bandbreite verglichen. Im Gegensatz zu den xu(15)-Testdaten liefert der HS(x) bei den realen Testdaten schlechte Resultate, da hier keine (“stückweise”) Gleichverteilung vorliegt. Dagegen liefert der Hybridselektivitätsschätzer in allen Fällen eine zum Teil deutliche Verbesserung gegenüber dem Kernselektivitätsschätzer. So reduziert sich der MRSF z.B. bei den rr2(22)-Testdaten von 19,9% beim KS-R(AOB-NS) auf 11,1% beim YS-R und kommt dabei dem MRSF des KS-R bei OGB von 10,8% recht nahe. Bei den ar2(18)-Testdaten ist der MRSF des YS-R von 14,9% sogar niedriger als der des KS-R(OGB) von 15,9%.

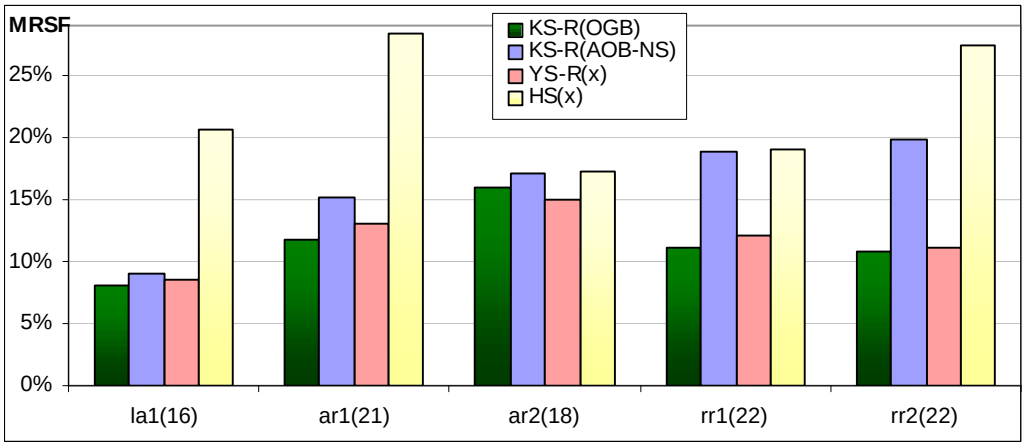


Abbildung 5.42: MRSF für den YS und verwandte Schätzer für reale Testdaten bei 1%-Anfragen.

In der nachstehenden Abbildung 5.43 wurden der MRSF des Hybridselektivitätsschätzers bei den realen Testdaten noch einmal mit den zuvor diskutierten Selektivitätsschätzern KS-R, EW-HS und ASHS jeweils mit geschätzter AOB verglichen. Dabei schneidet der Hybridselektivitätsschätzer teilweise deutlich besser ab als die anderen Selektivitätsschätzer. Lediglich bei den la1(16)-Testdaten ist der MRSF mit 8,0% etwas besser als beim YS-R mit 8,6%.

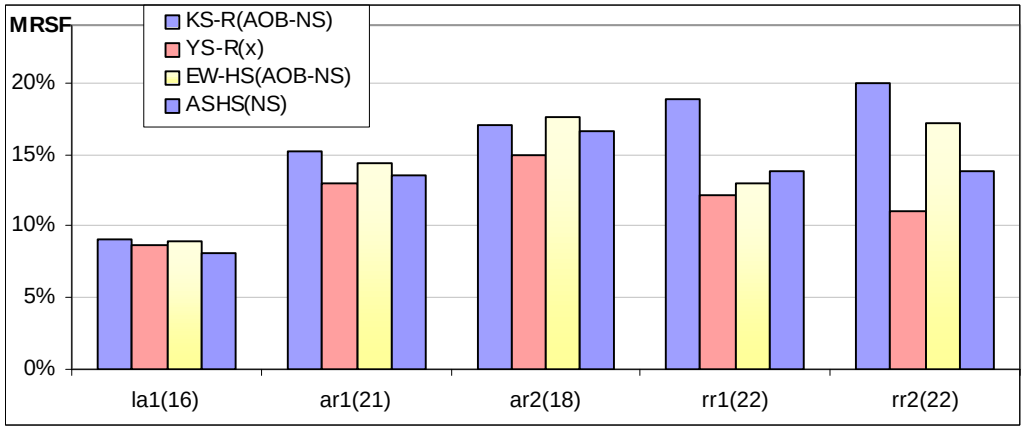


Abbildung 5.43: MRSF des KS-R, YS-R, EW-HS und ASHS jeweils mit AOB bei realen Testdaten.

Die Ergebnisse zeigen deutlich, daß im Falle der realen Testdaten das hybride Verfahren besser abschneiden kann als die anderen vorgestellten Selektivitätsschätzer. Daß dies nicht der Fall ist mit den künstlichen Testdaten zeigt die Notwendigkeit neben den theoretischen Untersuchungen die Verfahren zur Selektivitätsschätzung insbesondere anhand Experimenten mit realen Testdaten zu untersuchen, da solche in realen Datenbanken vorliegen. Sowohl der Kernselektivitätsschätzer als auch der Hybridselektivitätsschätzer wurden dabei in der Variante mit Randbehandlung untersucht. Zudem wurde die asymptotisch optimale Bandbreite bei allen Selektivitätsschätzern mit Hilfe der Normalskalierungsregel geschätzt. Sind geeignete Histogramme mit Bingrenzen an Sprungstellen bekannt, so kann der MRSF des Hybridselektivitätsschätzers mit AOB-NS sogar untern den MRSF des Kernselektivitätsschätzers mit optimaler gefundener Bandbreite fallen. Verbesserungen wären zudem durch die Verwendung von besseren Schätzern der AOB wie z.B. der direkten Plug-In-Regel beim Hybridselektivitätsschätzer zu erwarten. Fokus weiterer Forschungen in dieser Richtung sollte jedoch die Ablösung der hier angewendeten heuristischen Methode zur Festlegung der Bingrenzen durch automatische Sprungstellendetektoren, wie in Kapitel 3.5 andiskutiert, sein.

5.5.7 Abschließende Bewertung der Ergebnisse im univariaten Fall

Seien zum Abschluß von Abschnitt 5.5 die verschiedenen Ergebnisse der vorhergehenden Experimente vorgebracht. Dazu ist in Abbildung 5.44 der MRSF der vielversprechensten univariaten Selektivitätsschätzer bei 1%-Anfragen und diversen künstlichen und realen Testdaten

dargestellt. Als die vielversprechensten Verfahren haben sich bei den Untersuchungen die folgenden Verfahren herausgestellt:

- Equi-Width Histogrammselektivitätsschätzer mit Normalskalierungsregel: EW-HS(NS)
- Kernselektivitätsschätzer mit Randbehandlung und direkter Plug-In Regel 2. Stufe: KS-R(DPI2)
- Average Shifted Histogrammselektivitätsschätzer mit Normalskalierungsregel: ASHS(NS)
- Hybridselektivitätsschätzer mit manuell und heuristisch bestimmten Histogrammbins sowie Randbehandlung und Normalskalierungsregel: YS-R(x,NS)

Dabei wurde der Hybridselektivitätsschätzer lediglich bei den realen Testdaten angewendet, da die künstlichen Testdaten innerhalb der Domäne keine Sprungstellen aufweisen.

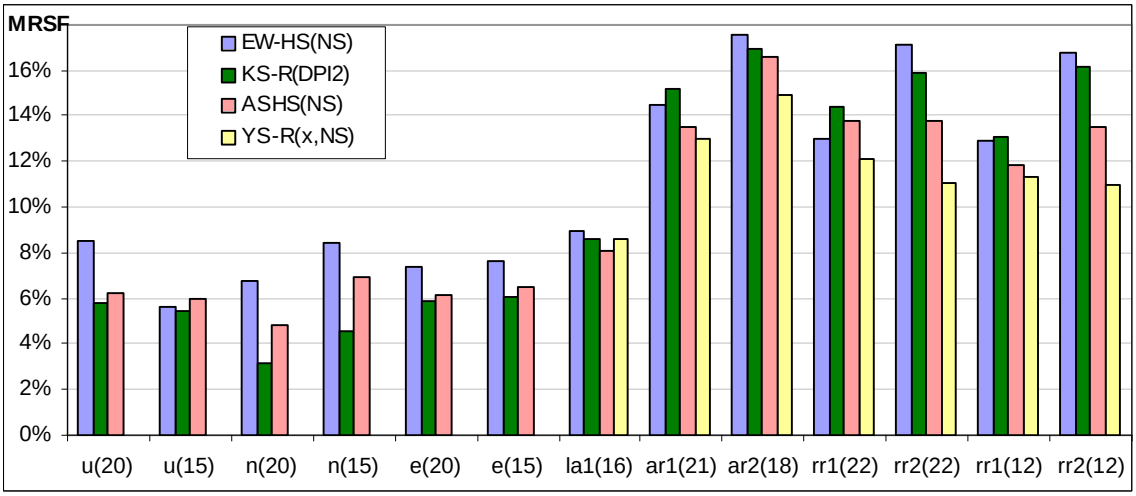


Abbildung 5.44: MRSF der vielversprechensten univariaten Selektivitätsschätzer.

Die Ergebnisse der künstlichen Testdaten zeigen in Abbildung 5.44, daß der Kernselektivitätsschätzer die besten Ergebnisse und der Histogrammselektivitätsschätzer die schlechtesten der drei Schätzer erzeugt. Hiervon abweichende Ergebnisse zeigen die Verfahren jedoch für die realen Testdaten. Hier erzeugen der Histogramm- und der Kernselektivitätsschätzer gleichermaßen schlechte Ergebnisse. Wie weiter oben gezeigt wurde, ist dies nicht allein durch eine schlechte Wahl der Bandbreite verursacht, da die jeweiligen Schätzer selbst mit optimaler gefundener Bandbreite nicht viel besser abschneiden. Eine teilweise sogar deutliche Verbesserung wird jedoch durch Verwendung des Hybridselektivitätsschätzers mit erreicht. Hier ist jedoch noch die Entwicklung von automatischen Verfahren zur Detektion von Sprungstellen und damit der Bildung der Histogrammbins erforderlich. Eine weitere Verbesserung des Hybridselektivitätsschätzers ist außerdem durch die Verwendung der DPI-2 Regel zur Bestimmung der AOB innerhalb der Histogrammbins zu erwarten.

5.6 Ergebnisse im bivariaten Fall

In diesem Kapitel werden die Ergebnisse der bivariaten Selektivitätsschätzung präsentiert. Dazu werden zunächst die beiden klassischen bivariaten Histogrammselektivitätsschätzer (Equi-Width und Equi-Depth) untereinander aber auch mit dem verwandten Verfahren des KD-Baum-Selektivitätsschätzers verglichen. Univariate Histogrammselektivitätsschätzer werden auf die auf eine eindimensionale Z-Ordnung abgebildeten Daten angewendet. Des weiteren werden der Average Shifted Histogrammselektivitätsschätzer und die Kernselektivitätsschätzer mit und ohne Randbetrachtung untersucht. Ein besonderes Augenmerk wird hier auf den Einfluß der Korrelation und der Möglichkeit der Ergebnisverbesserung durch Nutzung des Korrelationskoeffizienten bei der Bandbreitenbestimmung gelegt.

Die Ergebnisse der Selektivitätsschätzung mittels der Gleichverteilungsannahme (GS) und der direkten Selektivitätsschätzung (DS) sollen aufgrund der Einfachheit und weiten Verbreitung der Verfahren als Referenzwerte in den folgenden Vergleichen dienen. Deshalb werden die Ergebnisse mit diesen Verfahren im folgenden kurz vorgestellt und diskutiert. Die Verfahren wurden jeweils basierend auf einer Stichprobe von 2000 Elementen mit 1000 Testdaten der Größe 1%, 2%, 5% und 10% getestet.

5.6.1 Ergebnisse der Selektivitätsschätzung mittels Gleichverteilungsannahme

Da es sich bei der Gleichverteilungsannahme um ein parametrisches Verfahren handelt, sind gute Ergebnisse nur zu erwarten, wenn die wahre Verteilung der Daten näherungsweise zur Familie der Gleichverteilungen gehört. Dies ist bei den Testdatensätzen lediglich bei den gleichverteilten Datensätzen u(15x15) der Fall, wo der MRSF zwischen 0,6% (10%-Anfragen) und 2,5% (1%-Anfragen) liegt. Bei allen anderen Testdatensätzen sind die Ergebnisse nicht akzeptabel, insbesondere bei schiefen Verteilungen wie z.B. den exponential verteilten Datensätzen e(15x15), wo der MRSF sogar auf über 400% (10%-Anfragen) ansteigen kann. Abbildung 5.45 zeigt die unterschiedlichen Ergebnisse der GS bei den verschiedenen Verteilungen und einer Anfragegröße von 1%.

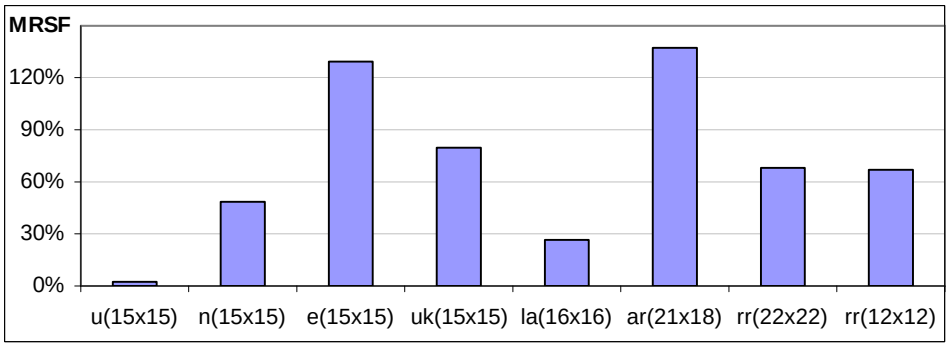


Abbildung 5.45: MRSF der Selektivitätsschätzung aufgrund der Gleichverteilungsannahme (GS) bei 1%-Anfragen.

5.6.2 Ergebnisse der direkten Selektivitätsschätzung

Die Ergebnisse der direkten Selektivitätsschätzung sind unabhängig von der zugrunde liegenden wahren Verteilung und zunächst als durchaus akzeptabel zu bezeichnen. Abbildung 5.46 zeigt die unterschiedlichen Ergebnisse der DS bei den verschiedenen Verteilungen und einer Anfragegröße von 1%. Der MRSF liegt hier zwischen 7,7% (uk15x15) und 18,0% (u15x15).

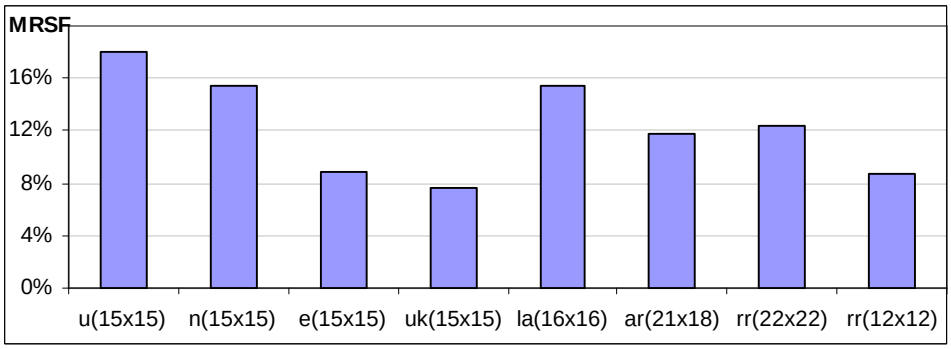


Abbildung 5.46: MRSF der direkten Selektivitätsschätzung (DS) bei 1%-Anfragen.

Es ist daher wie im univariaten Fall das Ziel der folgenden Experimente insbesondere die Überlegenheit der verfeinerten nicht-parametrischen Verfahren gegenüber der direkten Selektivitätsschätzung aufzuzeigen.

5.6.3 Ergebnisse der verschiedenen Histogrammselektivitätsschätzer

Im folgenden werden zunächst die Ergebnisse der verschiedenen Histogramm ähnlichen Selektivitätsschätzer miteinander verglichen. Dazu gehören der Equi-Width- und Equi-Depth-Histogrammselektivitätsschätzer (EW-HS, ED-HS), der KD-Baum-Selektivitätsschätzer (KDBS), der Average Shifted Histogrammselektivitätsschätzer (ASHS) und die univariaten Histogrammselektivitätsschätzer, die auf die auf eine eindimensionale Z-Ordnung abgebildeten Daten angewendet werden.

Zuerst wurden die verschiedenen Histogrammselektivitätsschätzer mit variierender Anzahl von Bins getestet und die jeweils besten Ergebnisse (mit optimaler gefundener Anzahl Bins OGB) miteinander verglichen. Man beachte, daß dabei für die verschiedenen Histogrammselektivitätsschätzer bei gleichen Testdaten unterschiedliche Binweiten bzw. eine unterschiedliche Anzahl von Bins vorliegen können.

Da die Gleichverteilungsannahme aus vorigem Abschnitt einem Histogrammschätzer mit einem einzigen Bin entspricht, ist bei den gleichverteilten Testdaten u(15x15) zu erwarten, daß die optimale Anzahl von Bins in diesem Falle ebenfalls - unabhängig vom jeweiligen Partitionierungsverfahren - einem einzigen Bin entspricht. Dies wird durch die Experimente bestätigt, wo beim GS das beste Ergebnis erzielt wurde (MRSF = 2,47%). Es zeigt sich, daß die Ergebnisse beim EW-HS durchweg besser als beim ED-HS sind, vgl. Abbildung 5.47 für 1%-Anfra-

gen. Nur bei den gleichverteilten Testdaten u(15x15) ist der ED-HS minimal besser (MRSF(EW-HS) = 4,09%, MRSF(ED-HS) = 4,01% jeweils bei k = 9 Bins). Der EW-HS ist somit bei metrischen Daten dem ED-HS vorzuziehen. Man beachte, daß die hier erzielten Ergebnisse von in anderen Veröffentlichungen präsentierten Ergebnissen abweichen ([Poosala et al. 96]).

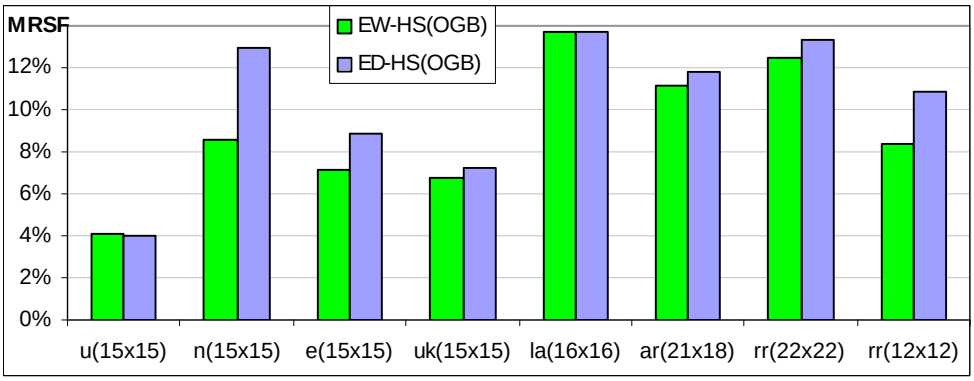


Abbildung 5.47: MRSF von EW-HS und ED-HS bei 1%-Anfragen.

Für das Equi-Width-Histogramm sind in Kapitel 4.4 Verfahren vorgestellt worden zur Schätzung einer asymptotisch optimalen Bandbreite (AOB). In Abbildung 5.48 ist daher die Qualität der Normalskalierungsregel im bivariaten Fall im Vergleich zu den Ergebnissen bei der OGB dargestellt. Dabei zeigt sich, daß die Normalskalierungsregel insbesondere bei den uk(15x15)- und den rr(*pxp*)-Testdaten sehr schlecht ist. Dies wird durch in Tabelle 5.12 durch die Angabe der Anzahl Bins bestätigt. Die Tabelle zeigt u.a., daß die durch die Normalskalierungsregel gewählte Anzahl Bins i.a. zu niedrig ist, was zu einer Überglättung der Schätzung führt. Wie man in Kapitel 5.6.5 sehen kann, liegt jedoch der Hauptgrund in der hohen Korrelation der Daten. Der Fehler kann durch ein verbessertes Normalskalierungsverfahren zum Teil deutlich verbessert werden, so daß dieses ein für die Praxis brauchbares Verfahren darstellt.

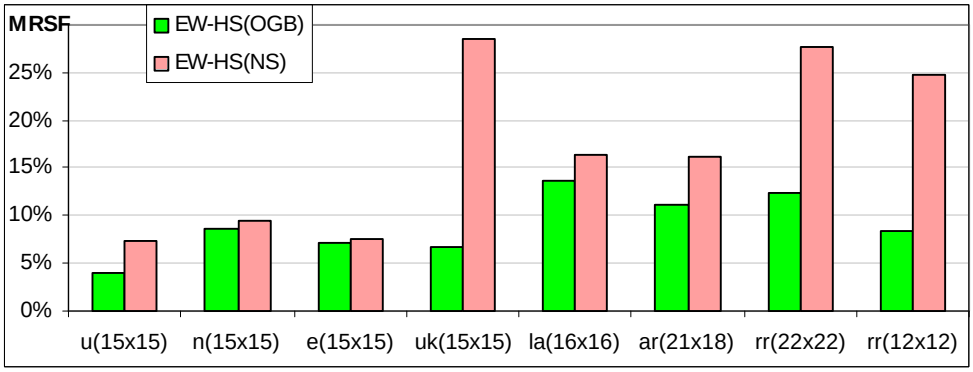


Abbildung 5.48: MRSF von EW-HS bei OGB und AOB-NS bei 1%-Anfragen.

Testdaten	u(15x15)	n(15x15)	e(15x15)	uk(15x15)	la(16x16)	ar(21x18)	rr(22x22)	rr(12x12)
OGB	9	81	1200	1600	195	1365	1974	1974
AOB-NS	49	100	1452	49	56	216	90	90

Tabelle 5.12: Anzahl Bins *k* bei OGB und AOB-NS für EW-HS.

Mit dem KD-Baum-Selektivitätsschätzer KDBS wurde in Kapitel 5.5 ein weiterer an Histogramme angelehnter Selektivitätsschätzer vorgestellt. Leider lassen sich durch den KDBS keine besseren Ergebnisse erzielen, wie die Experimente mit OGB zeigen. Stattdessen sind die Ergebnisse abgesehen leicht schlechter, vgl. Abbildung 5.49. Ein weiteres offenes Problem beim KDBS ist die Wahl der (möglichst) optimalen Anzahl Bins. Der KDBS wird daher im folgenden nicht weiter betrachtet.

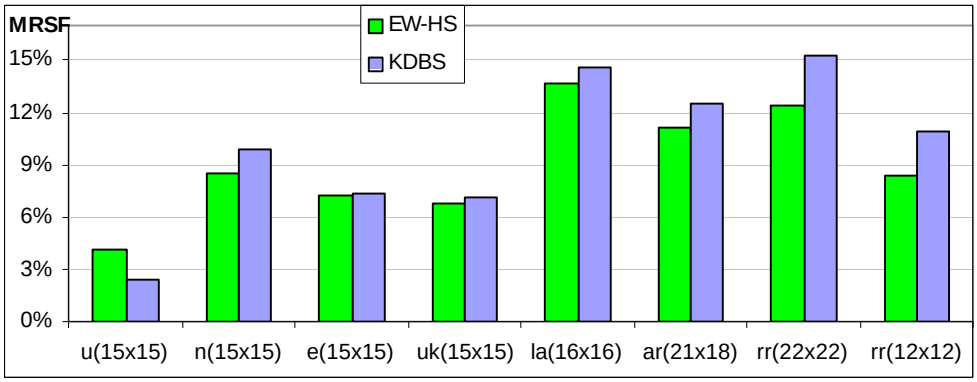


Abbildung 5.49: MRSF von EW-HS und KDBS bei OGB und 1%-Anfragen.

Als eine Möglichkeit univariate Histogrammselektivitätsschätzer auf multivariate Daten anzuwenden, wurde in Kapitel 3.4.2 ein Verfahren vorgestellt, das zunächst die multivariaten Daten auf eine eindimensionale Ordnung abbildet. Hierauf lassen sich unterschiedliche univariate HS anwenden (Equi-Width-, Equi-Depth, Max-Diff, Max-Diff-Z-). Dabei schneiden die diversen HS bei den verschiedenen Testdaten unterschiedlich ab, so daß keinem der vier Verfahren unbedingt der Vorzug gegeben werden kann, vgl. Abbildung 5.50, wo der MRSF bei 1%-Anfragen angewendet wurde. Die Anzahl Bins wurde dabei nach der Abbildung auf die Z-Ordnung durch die Normalskalierungsregel für EW-HS mit $d = 1$ bestimmt. Aufgrund des trotz Verwendung eines rekursiven Algorithmus zur Berechnung der Schnittmenge von Anfragebereich und Histogrammbins (vgl. Kapitel 3.4.2) sehr hohen Rechenaufwands war die Bestimmung einer OGB nicht möglich.

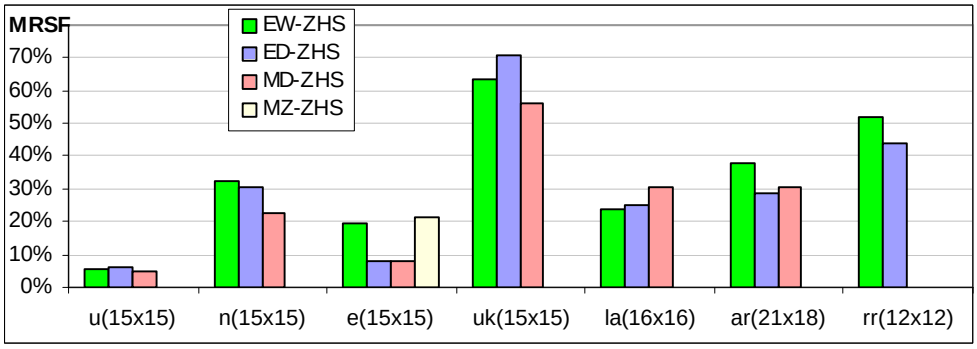


Abbildung 5.50: MRSF von verschiedenen ZHS bei 1%-Anfragen.

In Abbildung 5.51 wird daher der EW-ZHS mit dem EW-HS bei 1%-Anfragen verglichen. Die Bandbreite wurde in beiden Fällen mit der Normalskalierungsregel bestimmt. Die Grafik zeigt, daß der EW-ZHS mit Ausnahme der gleichverteilten Testdaten einen deutlich höheren MRSF erzeugt. Ein weiterer schwerwiegender Nachteil des EW-ZHS ist der stark erhöhte Rechenaufwand bei der Berechnung des Schnitts von Anfragebereich und univariatem Histogramm-Bin auf der Z-Kurve. Aus diesen beiden Gründen werden die ZHS im folgenden nicht weiter betrachtet - ihre Verwendung ist nicht empfehlenswert.

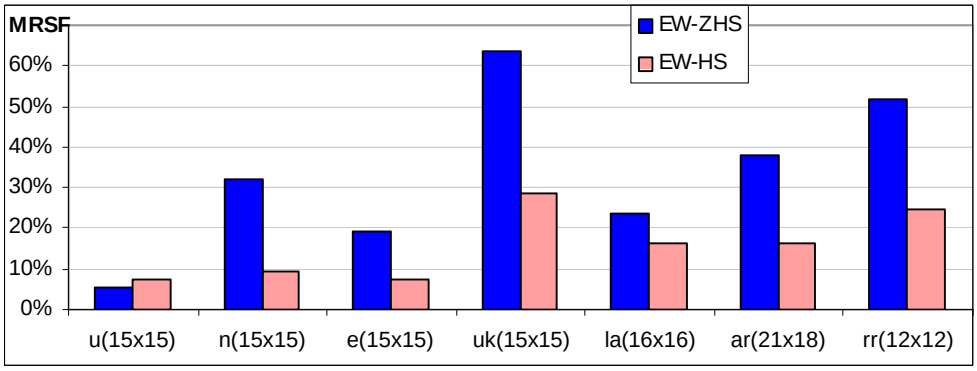


Abbildung 5.51: MRSF von EW-HS und EW-ZHS bei AOB-NS und 1%-Anfragen.

Bereits im univariaten Fall wurde der EW-HS mit dem Average Shifted Histogrammselektivitätsschätzer ASHS verglichen. Abbildung 5.52 zeigt die Ergebnisse der beiden Selektivitätsschätzer im bivariaten Fall zusammen mit den Ergebnissen des direkten Selektivitätsschätzers DS. Beim EW-HS und beim ASHS wurde in beiden Fällen die OGB betrachtet. Dabei zeigt sich, daß die Verwendung des ASHS keine Verbesserung bringt. Vielmehr ist der höhere Berechnungsaufwand des ASHS zu berücksichtigen, der im bivariaten Fall noch höher ausfällt als im univariaten Fall, vgl. Gleichung (3.20). Daher ist die Verwendung ASHS zur Selektivitätsschätzung nicht zu empfehlen. Weiterhin ist auffällig, daß der EW-HS(OGB) bei den realen

Testdaten meistens nur gering besser ist als der DS. Bei realen Testdaten ist daher der Aufwand bei der Verwendung eines Histogrammselektivitätsschätzers gegenüber dem direkten Selektivitätsschätzers kaum zu rechtfertigen.

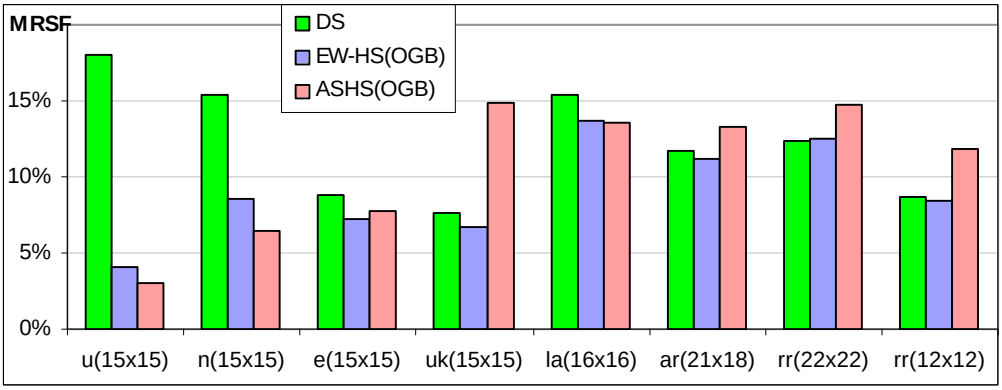


Abbildung 5.52: MRSF von DS, EW-HS und ASHS bei 1%-Anfragen und OGB.

Letztere Annahme wird weiterhin durch Abbildung 5.53 unterstützt, die den DS mit dem EW-HS mit AOB-NS vergleicht. Lediglich bei den künstlichen Testdaten u(15x15), n(15x15) und e(15x15) ist der MRSF des EW-HS niedriger als der MRSF des DS. Bei den uk(15x15)-Testdaten ist der hohe MRSF des EW-HS durch die hohe Korrelation der Daten bedingt. Dieser Aspekt wird in Abschnitt 5.6.5 näher betrachtet. Bei den realen Testdaten ist der MRSF des DS jedoch generell niedriger, so daß in diesem Fall der DS dem EW-HS(NS) vorzuziehen ist.

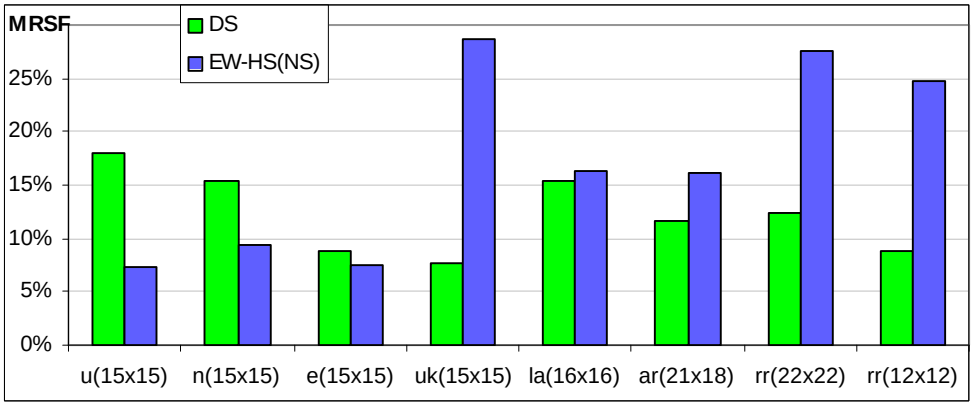


Abbildung 5.53: MRSF von DS und EW-HS bei 1%-Anfragen und AOB-NS.

Zusammenfassend läßt sich sagen, daß sich die Verwendung von weiteren Histogrammselektivitätsschätzern wie den hier vorgestellten gegenüber dem Equi-Width Histogrammselektivitätsschätzer in der Praxis nicht lohnt, da i.a. keine besseren Resultate zu erwarten sind. Hinzu kommt, daß außer für den EW-HS keine geeigneten Verfahren zur Bestimmung einer AOB

bekannt sind. Von der Verwendung der Z-Ordnung ist ganz Abstand zu nehmen, da die Ergebnisse wesentlich schlechter sind und der Berechnungsaufwand extrem hoch. Lediglich der Average Shifted Histogrammselektivitätsschätzer weist bessere Ergebnisse auf, hat aber einen höheren Berechnungsaufwand zur Folge. Erstaunlich ist das im Vergleich gute Abschneiden des direkten Selektivitätsschätzer. Im nächsten Abschnitt wird untersucht, inwieweit Kernselektivitätsschätzer eine Verbesserung in den Ergebnissen bringen.

5.6.4 Ergebnisse der verschiedenen Kernselektivitätsschätzer

In diesem Abschnitt werden die Kernselektivitätsschätzer untersucht und mit dem Histogrammselektivitätsschätzer und dem direkten Selektivitätsschätzer verglichen. Bereits im univariaten Fall wurde ausführlich die beim Kernschätzer auftretenden Randprobleme diskutiert. Abbildung 5.54 vergleicht den MRSF bei den KS ohne (KS-O) und mit Randbehandlung (Spiegelung KS-S u. Randkern KS-R) bei jeweils optimaler gefundener Bandbreite. Dabei zeigen sich allerdings nicht so drastische Unterschiede wie im univariaten Fall. Lediglich bei den gleichverteilten Testdaten u(15x15) zeigt sich eine deutlichere und bei den exponentiell verteilten Testdaten e(15x15) lediglich eine leichte Verbesserung. Hierbei spielt Wahl der optimalen Bandbreite sicherlich eine Rolle. So ergibt sich beim KS-O eine wesentliche geringere Bandbreite als bei den anderen beiden KS-Verfahren, so daß der Einfluß des Randfehlers wieder ausgeglichen wird.

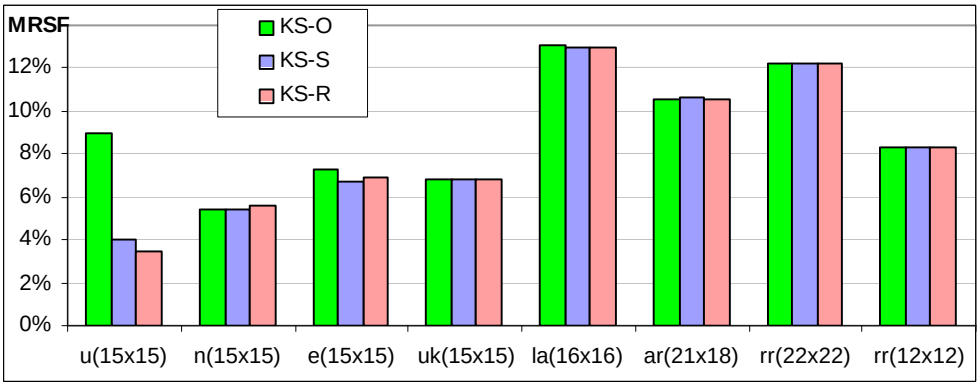


Abbildung 5.54: MRSF von KS-O mit KS-S und KS-R mit jeweils OGB bei 1%-Anfragen.

Abbildung 5.55 zeigt nun beim KS-S den Unterschied bei Verwendung der OGB und der mittels Normalskalierungsregel geschätzten AOB. Wie beim HS ist der MRSF des KS-S(AOB-NS) bei den realen Testdaten und insbesondere bei den uk(15x15)-Testdaten wesentlich schlechter als praktisch möglich. Dies wird in Tabelle 5.12 durch die Angabe der Bandbreite bestätigt. Die Tabelle zeigt u.a., daß die durch die Normalskalierungsregel gewählte Bandbreite i.a. zu hoch ist, was zu einer Überglättung der Schätzung führt. Auch hier kann, wie in Abschnitt 5.6.5 gezeigt wird, eine deutliche Verbesserung durch Berücksichtigung des Korrelationskoeffizienten bei der Schätzung der AOB erreicht werden.

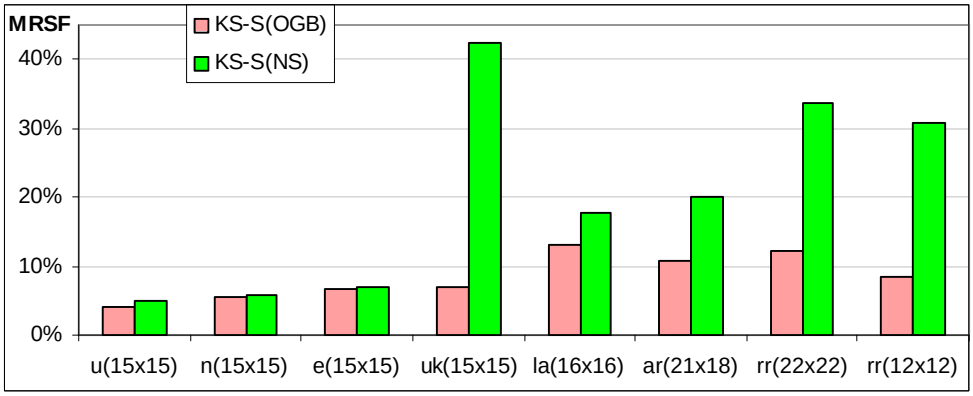


Abbildung 5.55: MRSF von KS-S mit OGB und AOB-NS bei 1%-Anfragen.

Testdaten	OGB [h_1, h_2]	AOB-NS [h_1, h_2]
u(15x15)	11.800	5.901
	11.800	5.892
n(15x15)	5.100	4.502
	5.100	4.498
e(15x15)	1.230	1.518
	943	1.161
uk(15x15)	795	5.901
	795	5.935
la(16x16)	4.410	11.709
	4.074	10.780
ar(21x18)	28.522	120.718
	10.478	44.713
rr(22x22)	46.791	575.568
	43.866	539.602
rr(12x12)	55	562
	50	527

Tabelle 5.13: Bandbreite *h* bei OGB und AOB-NS für EW-HS.

Abbildung 5.56 vergleicht den MRSF des ASHS, des KS-S und des DS bei 1%-Anfragen. Bei den künstlichen Testdaten ergeben sich für den KS-S Werte die sowohl besser als beim ASHS als auch beim DS sind. Anders sieht dies bei den realen Testdaten aus. Hier liefert der KS-S schlechtere Ergebnisse als die beiden anderen Verfahren.

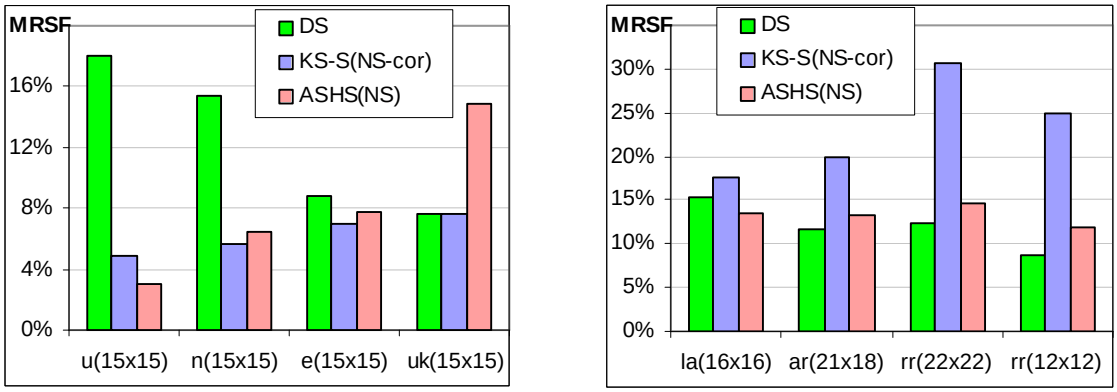


Abbildung 5.56: Vergleich des MRSF von DS, KS und ASHS bei 1%-Anfragen **a)** bei künstlichen Daten und **b)** bei realen Daten.

Um dies genauer zu untersuchen ist in Abbildung 5.57 der DS noch einmal mit dem KS-S bei OGB verglichen. Die Ergebnisse zeigen, daß der Kernselektivitätsschätzer bei allen Testdatensätzen bessere Resultate liefern kann. Allerdings dürfte es in realen Anwendungen schwierig sein, dieses Potential vollständig zu nutzen. Eine Verbesserung der Ergebnisse kann durch ausgefeiltere Verfahren zur Schätzung der AOB, wie sie im univariaten Fall in Kapitel 5.5.5 untersucht wurden, erreicht werden.

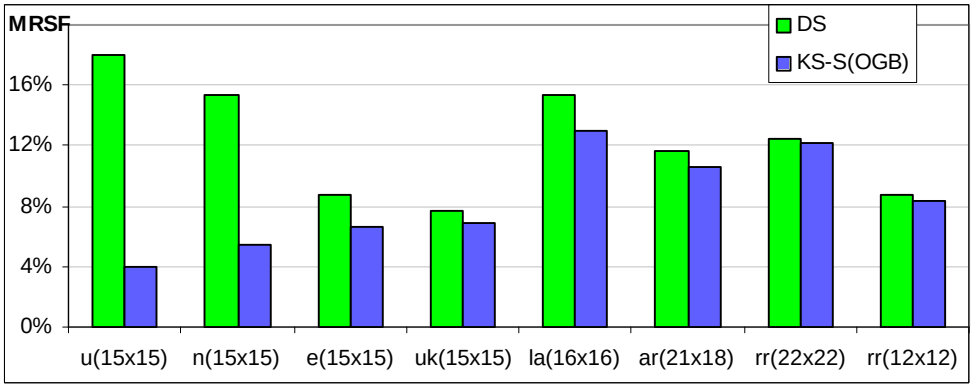


Abbildung 5.57: MRSF von DS und KS(OGB) bei 1%-Anfragen **a)** bei künstlichen Daten und **b)** bei realen Daten.

5.6.5 Einfluß der Korrelation auf die Selektivitätsschätzung

Im folgenden wird die Auswirkung von angenommenen Korrelationen auf den MRSF des EW-HS untersucht sowie die Verbesserungsmöglichkeiten bei Verwendung des Korrelationskoeffizienten in der Normalskalierungsregel (NS-Cor). In Tabelle 5.14 ist für die jeweiligen Testdaten der Korrelationskoeffizient (berechnet mit MS-Excel 97 und gerundet auf 4 Nachkommastellen) aufgeführt. Der Korrelationskoeffizient deutet eine extrem starke Korrelation für die uk()-Testdaten an, eine mittlere Korrelation für die rr()-Testdaten, eine leichte Korrelation für die

la()- und ar()-Testdaten sowie nahezu keine Korrelation für die u()-, n()- und e()-Testdaten. Entsprechend fallen die Ergebnisse in Abbildung 5.58 für 1%-Anfragen aus. Der MRSF ist bei Verwendung der Normalskalierungsregel ohne Korrelationskoeffizient bei den u-Testdaten extrem schlecht (28,6%), während die Verwendung des Korrelationskoeffizienten eine deutliche Verbesserung bringt (7,4%), so daß der MRSF nahezu dem der OGB entspricht (6,8%). Ebenso sind die MRSF bei den rr()-Testdaten sehr schlecht im Vergleich zum MRSF bei der OGB. Die Verwendung des Korrelationskoeffizienten bei der Normalskalierungsregel bringt auch hier eine wenn auch nicht ganz so deutliche Verbesserung. Bei den anderen Testdaten ergibt sich erwartungsgemäß keine Verbesserung durch Verwendung des Korrelationskoeffizienten bei der Normalskalierungsregel.

Testdaten	u(15x15)	n(15x15)	e(15x15)	uk(15x15)	la(16x16)	ar(21x18)	rr(22x22)	rr(12x12)
Korrelations-koeffizient	-0,002	0,007	-0,005	0,989	0,247	0,195	-0,449	-0,631

Tabelle 5.14: Korrelationskoeffizienten der Testdaten

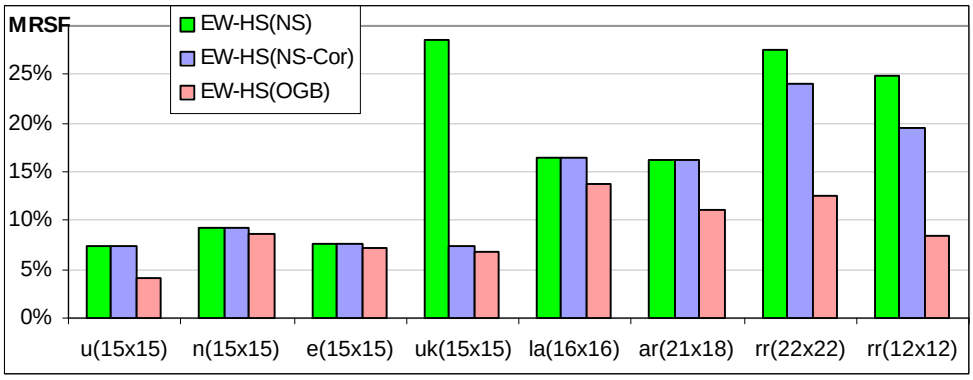


Abbildung 5.58: Auswirkung der Korrelation auf den MRSF von EW-HS mit unterschiedlich ermittelten Binweiten bei 1%-Anfragen.

Entsprechend zu den Untersuchungen zum EW-HS fallen die Ergebnisse beim KS-S wie in Abbildung 5.59 für 1%-Anfragen dargestellt aus. Der MRSF ist unter Verwendung der Normalskalierungsregel ohne Korrelationskoeffizient bei den u()-Testdaten extrem schlecht (42,3,6%), während die Verwendung des Korrelationskoeffizienten eine deutliche Verbesserung bringt (7,6%), so daß der MRSF nahezu dem der OGB entspricht (6,8%). Ebenso sind die MRSF bei den rr()-Testdaten sehr schlecht im Vergleich zum MRSF bei der OGB. Die Verwendung des Korrelationskoeffizienten bei der Normalskalierungsregel bringt auch hier eine Verbesserung. Bei den anderen Testdaten ergibt sich erwartungsgemäß keine Verbesserung durch Verwendung des Korrelationskoeffizienten bei der Normalskalierungsregel.

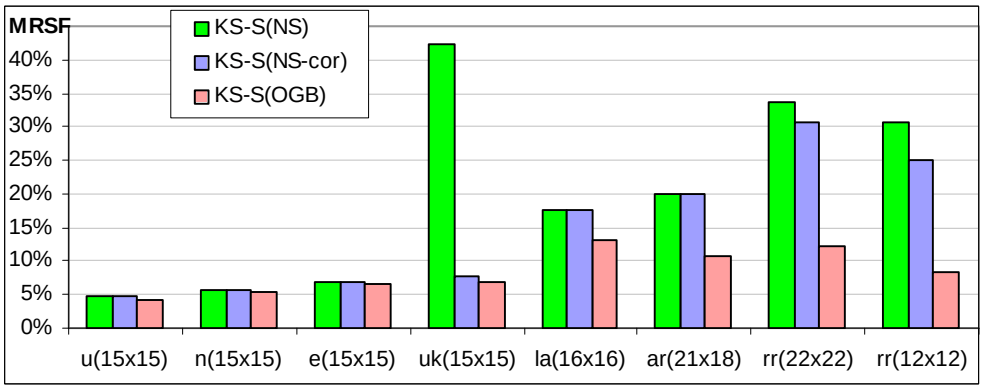


Abbildung 5.59: MRSF von KS-S bei 1%-Anfragen mit unterschiedlich ermittelten Bandbreiten.

5.6.6 Abschließende Bewertung der Ergebnisse im bivariaten Fall

Zum Abschluß dieses Kapitels seien die verschiedenen bivariaten Selektivitätsschätzer mit den vielversprechensten Varianten noch einmal in einen gemeinsamen Zusammenhang gestellt. Für jeden Testdatensatz werden die Ergebnisse der folgenden Verfahren berichtet:

- direkter Selektivitätsschätzer: DS
- Equi-Width Histogrammselektivitätsschätzer mit Normalskalierungsregel unter Berücksichtigung des Korrelationskoeffizienten: EW-HS(NS-cor)
- Average Shifted Histogrammselektivitätsschätzer mit Normalskalierungsregel: ASHS(NS)
- Kernselektivitätsschätzer mit Spiegelung und Normalskalierungsregel unter Berücksichtigung des Korrelationskoeffizienten: KS-S(NS-cor)

Abbildung 5.60 zeigt den MRSF dieser Verfahren bei 1%-Anfragen und verschiedenen zweidimensionalen Testdaten. Wie im univariaten Fall zeigt sich bei den künstlichen und den realen Testdaten ein unterschiedliches Bild. Bei den künstlichen Testdaten liefert der Kernselektivitätsschätzer gute Ergebnisse, wobei die anderen Verfahren teilweise gleich gute oder auch bessere Ergebnisse (ASHS bei u(15x15)-Testdaten) hervorbringen können. Anders sieht dies bei den realen Testdaten aus - hier gibt es bei den experimentellen Ergebnissen starke Abweichungen zu den theoretischen Ergebnissen. Der Kernselektivitätsschätzer liefert durchweg die schlechtesten Ergebnisse, während der direkte Selektivitätsschätzer insgesamt am besten abschneidet. Die Ursache ist in der fehlenden Voraussetzung der Glattheit der realen Testdaten zu suchen. Bei Anwendungen auf realen Datenbanken ist daher von den hier untersuchten Verfahren der direkte Selektivitätsschätzer vorzuziehen.

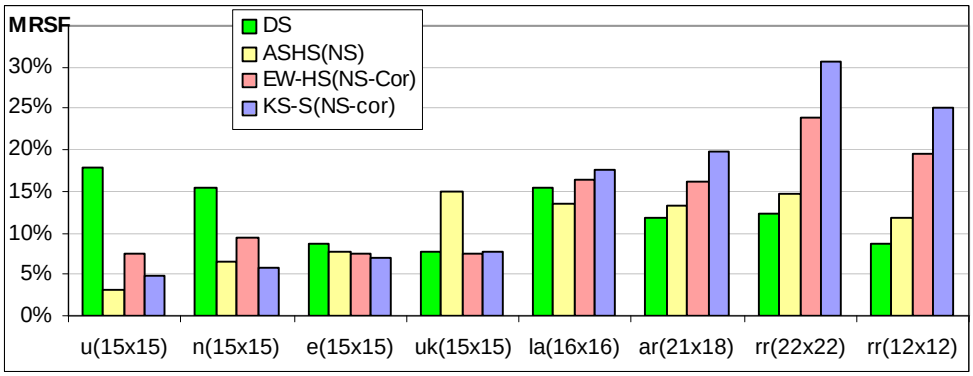


Abbildung 5.60: MRSF von DS, ASHS(NS), EW-HS(NS-cor) und KS-S(NS-cor) bei 1%-Anfragen.

Um das Potential der verschiedenen Verfahren zu beurteilen, ist in Abbildung 5.61 der MRSF für den EW-HS und den KS-S mit jeweils optimaler gefundener Bandbreite im Vergleich zum MRSF des DS dargestellt. Die Testdaten und Anfragen sind die gleichen wie in Abbildung 5.60. Hier zeigt sich, daß sowohl bei den künstlichen als auch bei den realen Testdaten der direkte Selektivitätsschätzer das schlechteste und der Kernselektivitätsschätzer das beste Ergebnis liefert, was eher der Theorie entspricht. Allerdings sind die Unterschiede bei den realen Testdaten nicht ganz so gravierend wie im univariaten Fall. Dies zeigt immerhin, daß durch eine bessere Schätzung der asymptotisch optimalen Bandbreite das oben bei realen Testdaten aufgezeichnete Bild durchaus verbessert werden könnte zugunsten der Kernselektivitätsschätzer. Hier besteht somit noch weiterer Forschungsbedarf.

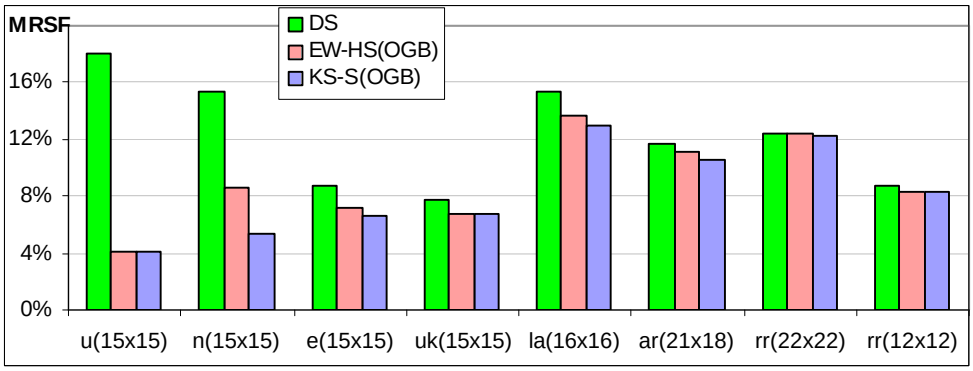


Abbildung 5.61: MRSF von EW-HS(OGB), KS-S(OGB) und DS bei 1%-Anfragen.

6. Diskussion und Ausblick

Selektivitätsschätzung ist eine wichtige Aufgabe von Datenbanksystemen und spielt insbesondere im Zuge des anwachsenden Datenvolumens solcher Systeme eine unerläßliche Rolle. Dabei gewinnt auch verstärkt die multivariate Selektivitätsschätzung an Bedeutung.

Das Problem der Selektivitätsschätzung wurde in einer Reihe von Veröffentlichungen betrachtet ([Muralikrishna & DeWitt 88], [Lipton & Naughton 90], [Whang et al. 94], [Chen & Roussopoulos 94], [Ioannidis & Poosala 95], [Poosala et al. 96], [Poosala & Ioannidis 97], [Matias et al. 98]). Bei den dort vorgestellten Verfahren handelt es sich im wesentlichen um nicht-parametrische Verfahren, insbesondere Histogrammschätzern und direkter Selektivitätsschätzung anhand der Stichprobe (Sampling). Dieser Bezug zur mathematischen Statistik wird jedoch erstmals in der vorliegenden Arbeit klar herausgestellt. Grundlage dazu ist die hier aufgestellte Beziehung zwischen der Dichte einer Datenverteilung und der Selektivität einer Anfrage auf einer dieser Datenverteilung entsprechenden Datenbank. Die Relation-Selektivität läßt sich dabei elegant über die Dichtefunktion beschreiben (Gleichung (2.5) in Kapitel 2.1). Dies erlaubt nun die eindeutige Übertragung von Verfahren der mathematischen Statistik auf die Problematik der Selektivitätsschätzung. Da der Typ der einer Datenbank zugrundeliegenden Verteilungsfamilie in realen Datenbanksystemen im allgemeinen nicht bekannt ist, ist die Verwendung von parametrischen Verfahren nicht sinnvoll und wurde daher in dieser Arbeit nicht weiter betrachtet. Die Konzentration auf nicht-parametrische Verfahren deckt sich mit den in den oben angegebenen Veröffentlichungen diskutierten Verfahren. Lediglich die Gleichverteilungsannahme wurde ebenfalls zu Vergleichszwecken in den Experimenten untersucht.

In den oben angegebenen Literaturstellen werden im allgemeinen wenig Angaben über die Art der in den Datenbank gespeicherten Daten gemacht. Diesbezüglich ist jedoch eine Unterscheidung zu treffen, da zum Beispiel die Schätzung von kategorialen Daten eine völlig andere Vorgehensweise erfordern kann als von metrischen Daten. Im Unterschied zu vorherigen Publikationen wurden daher in der vorliegenden Arbeit ausschließlich metrische Daten mit einer großen Domäne betrachtet.

Ein in anderen Veröffentlichungen im Zusammenhang mit Selektivitätsschätzern oft diskutiertes Problem ist der Hauptspeicherbedarf (z.B. [Ioannidis & Poosala 95]). So benötigt z.B. ein Histogramm mit 20 Bins sicherlich weniger Hauptspeicher (im univariaten Fall etwa 240 Byte bei 4 Byte pro Real-Zahl) als eine Stichprobe der Größe 2000 (im univariaten Fall etwa 8000 Byte bei 4 Byte pro Real-Zahl). Diese Diskussion ist jedoch historisch zu sehen. In heutigen Zeiten, in denen von Terabyte-Datenbanken die Rede ist und Hauptspeicher im GByte-Bereich bei großen Datenbanksystemen durchaus üblich sind, spielt es eine untergeordnete Rolle, ob ein Verfahren einige kByte mehr oder weniger benötigt.

Ein in der Literatur zur Selektivitätsschätzung häufig verwendetes Vorgehen besteht darin, als Stichprobe zur Generierung von Histogrammen die gesamte Datenbank zu verwenden. Dieser Aufwand ist jedoch bei großen Datenbanksystemen nicht realistisch und auch nicht notwendig. Stattdessen ist es für praktische Zwecke im allgemeinen ausreichend, eine genügend große

Stichprobe aus der Datenbank zu ziehen, vgl. dazu auch [Poosala et al. 96]. In dieser Arbeit wurde daher davon ausgegangen, daß immer eine repräsentative Stichprobe aus der Datenbank vorliegt.

Ein ideales Verfahren zur Selektivitätsschätzung, das auf alle Problemfälle gleichmaßen anwendbar ist, gibt es bisher nicht. Diese Arbeit zeigt dafür deutlich, welche Verfahren der nicht-parametrischen Statistik anwendbar sind und wo deren Vor- und Nachteile liegen. Dazu gehören Verfahren zur Selektivitätsschätzung auf Basis der empirischen Verteilungsfunktion sowie die Selektivitätsschätzung mittels Histogrammen, Average Shifted Histogrammen oder Kernfunktionen (Kapitel 3). Die Selektivitätsschätzung mittels Histogrammen wird bereits in den meisten Veröffentlichungen aus dem Bereich der Datenbanksysteme betrachtet. Die vorliegende Arbeit räumt aber mit dem dort herrschenden Vorurteil auf, daß dies das einzige nicht-parametrische Verfahren sei. Vielmehr handelt es sich dabei um **einen** nicht-parametrischen Ansatz von vielen. Die in dieser Arbeit vorgestellten Selektivitätsschätzer mit Average Shifted Histogrammen und Kernfunktion sind hingegen völlig neue Verfahren zur Selektivitätsschätzung. Insbesondere bei den Kernschätzern handelt es sich um ein Verfahren, das auch in der mathematischen Statistik in den letzten Jahren vermehrt an Bedeutung gewonnen hat ([Wand & Jones 95]).

In den oben zitierten Arbeiten zur nicht-parametrischen Selektivitätsschätzung mit Histogramm-basierten Schätzern ist die Anzahl der Histogrammbins als gegeben vorausgesetzt. Damit ist die Anzahl der Histogrammbins willkürlich gesetzt. Auch reale Datenbanksysteme bieten diesbezüglich keine Hilfestellung an [Oracle 99]. Tatsache ist aber, daß die Güte der Schätzung in hohem Maße von der verwendeten Anzahl von Bins abhängt - im uni- wie im multivariaten Fall (Kapitel 4 und 5). Das gleiche Problem taucht bei Average Shifted Histogrammschätzern sowie bei Kernschätzern mit der zu wählenden Bandbreite auf. In dieser Arbeit wird erstmalig im Zusammenhang mit der nicht-parametrischen Selektivitätsschätzung auf das Problem der Bandbreitenwahl eingegangen. Schätzungen einer asymptotisch optimalen Bandbreite lassen sich mit Hilfe des asymptotischen mittleren integrierten quadratischen Fehlers (AMISE) herleiten (Kapitel 4). Dabei kommen auch Verfahren zur adaptiven Wahl der Bandbreite zum tragen. Der AMISE erlaubt weiterhin die Angabe einer Konvergenzordnung für die verschiedenen Schätzverfahren (Kapitel 4.5).

Bei der Übertragung nicht-parametrischer Verfahren auf die Selektivitätsschätzung treten eine Reihe von Problemen auf und müssen gelöst werden. Dazu gehören u.a. das Randproblem bei Kernschätzern, die Integration von Kernfunktionen zur Selektivitätsschätzung, die Anwendung verschiedener Verfahren zur Bandbreitenbestimmung und die fehlende Glattheit bei realen Daten. Die Komplexität der Probleme nimmt dabei bei steigender Dimensionalität der Daten zu.

Bei der Randbehandlung von Kernschätzern zeigten sich in Bezug auf die experimentellen Ergebnisse die Spiegelung und die Benutzung von speziellen Randkernen als nahezu gleichwertig. Für die Spiegelung erwies sich der Randkern von [Dong & Simonoff 94] als gute und einzig realisierbare Möglichkeit zur praktischen Verwendung in einem Randkernselektivitätsschätzer (Kapitel 3.4.2).

Die Wahl der Bandbreite hat einen entscheidenden Einfluß auf die Güte der Histogramm- oder Selektivitätsschätzer. Unter Nutzung des AMISE stehen einige Regeln zur Schätzung einer asymptotisch optimalen Bandbreite zur Verfügung (Kapitel 4). Dabei erwies sich die Normalskalierungsregel bei Verteilungen, die nicht zur Familie der Normalverteilungen gehören, als unbefriedigend. Eine Verbesserung konnte nur durch Verwendung der direkten Plug-In Regel zweiter Ordnung erzielt werden. Adaptive Verfahren erschienen zunächst vielversprechend, scheitern allerdings in der Praxis oft an den Randproblemen, da bei variabler Bandbreite die oben vorgestellten Randbehandlungen nicht anwendbar sind. Da die geschätzten asymptotisch optimalen Bandbreiten teilweise trotzdem noch erheblich von der durch Ausprobieren aller möglichen Bandbreiten in den Experimenten gefundenen optimalen Bandbreite abweicht, besteht hier weiterer Forschungsbedarf.

Die verschiedenen hier diskutierten Verfahren wurden vor dem Hintergrund untersucht, diese einmal in realen Datenbanksystemen einsetzen zu können. Insofern waren die Schätzverfahren insbesondere mit realen Testdaten zu validieren. Die theoretischen Herleitungen der Selektivitätsschätzer wurden in dieser Arbeit durch ausführliche Experimente sowohl mit künstlich erzeugten als auch mit Daten aus realen Datenbanken (Geodaten) im uni- und bivariaten Fall validiert. Während die Selektivitätsschätzer bei den künstlichen Testdaten i.a. die aus der Theorie abgeleiteten Ergebnisse bestätigten, kam es bei den realen Testdaten doch zu erheblichen Abweichungen. So brachten die Kernselektivitätsschätzer bei den univariaten realen Testdaten keinen nennenswerten Gewinn gegenüber den Histogrammselektivitätsschätzern (Kapitel 5.5.4). Der Grund dafür liegt in der bei den realen Testdaten fehlenden Voraussetzung der Glattheit der zugrundeliegenden Verteilung, einer der wichtigen Voraussetzungen für die Anwendung von Kernschätzverfahren. Hier wurde im univariaten Fall eine Lösung unter Verwendung des Hybridselektivitätsschätzers gefunden (Kapitel 3.5 und 5.5.6), der eine Partitionierung des Datenraums entsprechend den “change points” der Dichtefunktion vornimmt und innerhalb der Histogrammbins die lokale Selektivität mittels Randkernen schätzt. Dabei kommen i.a. in den verschiedenen Histogrammbins auch (lokal) unterschiedliche Bandbreiten zum Tragen. Problematisch ist bei diesem Verfahren die Bestimmung der “change points”. Die in den Experimenten verwendete heuristische Methode ist sicherlich für praktische Zwecke nicht anwendbar, so daß auch hier Anlaß für weitere Forschungen besteht. Ein möglicher Ansatz besteht darin, Sprungstellen durch Verwendung “rechts-” und “linksseitiger” Kernschätzer zu detektieren, vgl. [Qiu 97] für solche Ansätze. Ein weiteres offenes Problem besteht in der Erweiterung des Hybridschätzer auf multivariate Daten.

Die direkte Selektivitätsschätzung auf Basis der Stichprobe ist ein einfaches nicht-parametrisches Verfahren, das brauchbare Ergebnisse liefert. Allerdings hat dieses Verfahren eine schlechtere Konvergenzordnung als die anderen diskutierten nicht-parametrischen Selektivitätsschätzer und liefert insbesondere im univariaten Fall wesentlich schlechtere Ergebnisse. Da dieser Unterschied bereits im bivariaten Fall schrumpft, ist die direkte Selektivitätsschätzung jedoch zur Zeit für höhervariate Selektivitätsschätzungen auch aufgrund der einfachen Implementierung eine praktische Alternative.

Histogrammschätzer sind ein klassisches nicht-parametrisches Schätzverfahren und werden auch im Bereich der Selektivitätsschätzung eingesetzt (z.B. Oracle Version 8.x). Es gibt zahlreiche Veröffentlichungen zu diesem Thema ([Muralikrishna & DeWitt 88], [Ioannidis & Poosala 95], [Poosala et al. 96], ...). Die verschiedenen Histogrammschätzer unterscheiden sich in der Art der Partitionierung. Zu den üblicherweise verwendeten gehören der Equi-Width- (mit konstanter Bandbreite) und der Equi-Depth-Histogrammschätzer. Diese wurden im uni- wie im bivariaten Fall experimentell untersucht. Dabei erwiesen sich die beiden Verfahren als nahezu gleichwertig mit im i.a. leicht besseren Ergebnissen beim Equi-Width-Histogrammselektivitätsschätzer. Im univariaten Fall wurde weiterhin das Max-Diff-Verfahren implementiert, welches jedoch i.a. deutlich schlechtere Ergebnisse liefert. Vorteil des Equi-Width-Verfahrens ist weiterhin das Vorliegen von geeigneten Verfahren zur Schätzung der asymptotisch optimalen Bandbreite. In dieser Bewertung unterscheidet sich die vorliegende Arbeit von anderen Veröffentlichungen ([Poosala et al. 96], [Muralikrishna & DeWitt 88]), in denen der Equi-Depth- oder der Max-Diff-Histogrammschätzer bevorzugt wird. Die unterschiedlichen Ergebnisse sind durch den Fokus dieser Arbeit auf metrische Daten begründet.

Interessant ist weiterhin die Verbindung von Histogrammschätzern mit geeigneten (multivariaten) Indexstrukturen für Datenbanksysteme. Dabei werden Indexverfahren verwendet, die den Datenraum vollständig und disjunkt in Partitionen aufteilen. Diese dienen als Grundlage für Histogramme. Beispiele solcher verwendeter Indexverfahren sind das Multi-Level-Grid-File ([Whang et al. 94]) und der KD-Baum ([Buskamp 97]). Durch die hierarchische Datenstruktur dieser Indexverfahren ist einerseits ein schnellerer Zugriff auf die Selektivität der Bins gegeben. Andererseits besteht die Möglichkeit, die Selektivitätsschätzung an verschiedenen Levels der Datenstruktur vorzunehmen, z.B. in Abhängigkeit von der Größe der Anfrage oder von der lokalen Datenverteilung. Allerdings gibt es keine Handhabe für die Anzahl der Bins bzw. das Level der Partitionierung. Inwieweit die Partitionierung letztendlich dem Optimum nahekommt, ist daher fraglich.

Eine weitere Alternative besteht in der Partitionierung mit Voronoi-Regionen. Der dabei auftretenden Überlappung der Voronoi-Regionen an den Rändern kann begegnet werden. Um den Rahmen dieser Arbeit nicht zu sprengen, wurde der Voronoi-Selektivitätsschätzer in dieser Arbeit nur beschrieben, aber nicht weiter implementiert. Erste Ergebnisse sind jedoch ermutigend für weitere Untersuchungen [Schneider 97].

Um univariate Histogrammselektivitätsschätzer direkt auf multivariate Daten anwenden zu können, lassen sich letztere mittels einer geeigneten Ordnung auf Basis einer raumfüllenden Kurve (hier Z-Kurve) auf univariate Daten abgebildet. Voraussetzung ist dafür die vorherige Diskretisierung der Daten. Allerdings läßt sich dieses Verfahren im multivariaten Fall nur für sehr kleine Domänen anwenden, da der Aufwand bei der Berechnung der Schnitte der Anfragebereiche mit den Histogramm-Bins trotz Anwendung eines rekursiven Algorithmus sehr hoch werden kann. Dieses Vorgehen ist somit zur Anwendung in der Praxis ungeeignet.

Allgemeiner Vorteil von Histogrammschätzern ist, daß sie einfach zu implementieren sind, nur wenig Informationen gespeichert werden müssen (Bingrenzen und Selektivität der Bins) und

die Selektivitätsschätzung schnell berechnet werden kann. Dazu müssen allerdings die Histogramm-Bins vorberechnet und ständig aktualisiert werden - ein unter Umständen mühsamer Vorgang. Auch können sich bei einer Veränderung der Verteilung nicht nur die Selektivitäten sondern auch die Bingrenzen ändern. Weiterere Nachteile von Histogrammschätzern sind die Abhängigkeit der Güte vom Anfangspunkt sowie die Tatsache, daß Histogrammschätzer i.a. im Gegensatz zur den Daten zugrundeliegenden wahren Funktion keine glatte Schätzfunktion liefern. Von der Theorie her haben Histogrammschätzer eine geringere Konvergenzordnung als einige andere nicht-parametrische Schätzverfahren (s.u.). Bei den Experimenten zeigten sie ebenfalls schlechtere Ergebnisse.

Eine Verfeinerung des Histogrammschätzers ist durch den Average Shifted Histogrammschätzer gegeben. Die starke Abhängigkeit des Schätzers vom Startpunkt entfällt. Der Average Shifted Histogrammschätzer wurde in dieser Arbeit erstmalig zur Selektivitätsschätzung verwendet. Er hat eine bessere Konvergenzordnung als der Histogrammschätzer und liefert sowohl im univariaten als auch im bivariaten Fall bessere Ergebnisse. Deshalb ist zumindest im uni- und bivariaten Fall der Average Shifted Histogrammschätzer dem einfachen Histogrammschätzer vorzuziehen. Allerdings ist der deutlich höhere Aufwand bei der Bestimmung der asymptotisch optimalen Bandbreite und bei der Implementierung zu berücksichtigen.

Ein weiteres völlig neues Vorgehen bei der Schätzung der Selektivität besteht in der Verwendung von Kernschätzern. Hierbei ist jedoch unbedingt die Behandlung von Randproblemen zu beachten - insofern nicht bekannt ist, daß am Rand keine Daten liegen. Kernschätzer zeigen sich gegenüber den vorherig besprochenen Schätzern durch ihre Glattheit und der Fähigkeit, lokale Strukturen besser abzubilden, aus. Bei realen Testdaten, bei denen die Voraussetzungen der Glattheit der zugrundeliegenden Verteilung nicht erfüllt sein muß, führt dies aber zu Problemen. So zeigen die Experimente bei realen Testdaten erheblich schlechtere Ergebnisse im Vergleich zu den Histogrammselektivitätsschätzern. Im Vergleich dazu schneiden die Kernselektivitätsschätzer (mit Randbehandlung) bei den künstlich erzeugten Testdaten mit glatter Verteilung besser ab. Ein weiterer Vorteil des Kernselektivitätsschätzers gegenüber den Histogrammselektivitätsschätzern ist, daß für die Schätzung a priori keine weiteren Informationen im Form von Histogrammbins und deren Selektivitäten berechnet werden müssen.

Um eine Lösung für das zuvor bei den Kernselektivitätsschätzern beschriebene Problem mit der fehlenden Glattheit bei realen Daten zu finden, wurde in dieser Arbeit für den univariaten Fall der Hybridselektivitätsschätzer entwickelt. Hierbei handelt es sich um eine Kombination von Histogramm- und Kernselektivitätsschätzer, bei der der Datenraum entsprechend der “Sprungstellen” der Verteilung partitioniert wird und die Annahme der Gleichverteilung innerhalb der Histogrammbins fallengelassen wird. Bei bekannten Sprungstellen ergaben die Experimente sehr gute Ergebnisse, die bei den realen Testdaten teilweise wesentlich besser als die anderen untersuchten Verfahren waren. Ein in diesem Zusammenhang für praktische Zwecke ungelöstes Problem ist die Detektion der Sprungstellen bzw. die Bestimmung geeigneter Bingrenzen. Hier besteht intensiver Forschungsbedarf ebenso wie bei der Erweiterung des Hybridselektivitätsschätzers auf multivariate Daten (s.o.).

Ein wichtiger Aspekt in der vorliegenden Arbeit war die Bestimmung einer geeigneten Bandbreite für die nicht-parametrischen Schätzverfahren. In der statistischen Literatur wird hierzu i.a. die asymptotisch optimale Bandbreite verwendet, die über die Minimierung des AMISE definiert ist. Zur Schätzung der asymptotisch optimalen Bandbreite existieren verschiedene Verfahren, von denen in dieser Arbeit hauptsächlich die Normalskalierungsregel, die direkte Plug-In Regel 2. Stufe und ein adaptives Verfahren betrachtet wurden. Diese wurden im Zusammenhang mit dem Kernselektivitätsschätzer als dem theoretisch vielversprechensten Schätzverfahren anhand univariater Testdaten experimentell untersucht. Sie lieferten alle relativ gute Ergebnisse, wobei die durch die direkte Plug-In Regel 2. Stufe bestimmte Bandbreite i.a. der gefundenen optimalen Bandbreite am nächsten kommt. Das adaptive Verfahren brachte keine nennenswerte Verbesserung, insbesondere da eine Randbehandlung bei variabler Bandbreite nicht ohne weiteres möglich ist. Es ist anzunehmen, daß die Ergebnisse bei den anderen diskutierten Selektivitätsschätzern ähnlich liegen, da die Normalskalierungsregel beim Histogrammselektivitätsschätzer und beim Average Shifted Histogrammselektivitätsschätzer ähnlich gelagerte Abweichungen von der gefundenen optimalen Bandbreite liefert. Gegebenenfalls ist dies durch entsprechende Experimente zu verifizieren. Bei der Verwendung des univariaten Kernselektivitätsschätzers ist in jedem Fall die direkte Plug-In Regel 2. Stufe zu empfehlen. Inwieweit sich die Ergebnisse auf multivariate Kernselektivitätsschätzer übertragen lassen, ist ebenfalls Gegenstand weiterer Forschungen auf diesem Gebiet.

Sind über die den Daten zugrundeliegende Verteilung weitere statistische Kenngrößen bekannt, so lassen sich diese Informationen gegebenenfalls zur Verbesserung der Verfahren nutzen. In [Scott 92] werden als Kandidaten hierfür die Schiefe, Kurtosis und im bivariaten Fall weiterhin der Korrelationskoeffizient diskutiert und der AMISE unter Ausnutzung der Kenngrößen entsprechend angepaßt. Entsprechend ändern sich die Gleichungen für die Bestimmung der asymptotisch optimalen Bandbreite. Dies wurden in der vorliegenden Arbeit für weitere Experimente zur Selektivitätsschätzung implementiert und brachte deutliche Verbesserungen sowohl bei der Kernselektivitätsschätzung als auch beim Histogrammselektivitätsschätzer.

Anhand der Ergebnisse in dieser Arbeit ist es schwierig eine allgemeine Empfehlung für die Verwendung des besten Selektivitätsschätzers für alle Problemfälle zu treffen, so daß dies einer differenzierteren Betrachtungsweise bedarf. Ist die Voraussetzung der Glattheit der den Daten zugrundeliegenden Verteilung gegeben, so sind sicherlich die Kernselektivitätsschätzer die geeignetste Wahl. Dabei ist zu beachten, daß bei Verteilungen mit Masse am Rand dies durch eine entsprechende Randbehandlung berücksichtigt wird. Leider ist bei Datenbanksystemen mit realen Daten nicht damit zu rechnen, daß die Voraussetzung der Glattheit der Verteilung gegeben ist. Existieren Sprungstellen der Verteilung und sind diese bekannt, so ist der Hybridselektivitätsschätzer die optimale Wahl. Solange praktikable Verfahren zur Detektion der Sprungstellen fehlen, ist in den Fällen, wo die Sprungstellen nicht bekannt sind, der Average Shifted Histogrammselektivitätsschätzer vorzuziehen. Dieser ist auch bei bivariaten Daten der zu empfehlende Selektivitätsschätzer. Die Untersuchung höherer Dimensionen steht noch aus und ist Gegenstand weiterer Forschungen auf diesem Gebiet.

Zusätzliche Verbesserungen der Selektivitätsschätzung in Datenbanksystemen lassen sich möglicherweise durch die Verwendung von Verfahren erreichen, die die Ergebnisse vorheriger Anfragen berücksichtigen und aus diesen lernen. Hierzu gehören Verfahren der anfrageabhängigen Selektivitätsschätzung ([Chen & Roussopoulos 94]) sowie die Verwendung adaptiver Verfahren wie z.B. künstlicher neuronaler Netze ([Rojas 93]). In diesem Zusammenhang ist auf die Beziehung von gewissen künstlichen neuronalen Netzen und Methoden der mathematischen Statistik hingewiesen. So werden in [Sarle 94] zum Beispiel mehrschichtige Perzeptronnetze (Multilayer-Perzeptron, MLP) mit Kernschätzverfahren verglichen. Allerdings ist bei der Verwendung künstlicher neuronaler Netze zu berücksichtigen, daß bei einer automatischen online-Selektivitätsschätzung der Speicher- und Rechenaufwand nicht zu groß werden darf.

Weitere Forschungsbereiche im Zusammenhang mit der Selektivitätsschätzung wäre die Erweiterung auf allgemeinere Anfragetypen wie z.B. Join-Anfragen von mehreren Datentabellen oder aggregierte Anfragen ([Hellerstein et al. 97]), sowie die Selektivitätsschätzung bei Datenbanksystemen mit räumlich ausgedehnten Objekten ([Samet 90]).

Anhang A

A.1 Notation

Tabelle A.1 gibt einen Überblick über die in dieser Arbeit verwendete Notation.

d	Dimension
R	Relation
I	Instanz einer Relation
N	Anzahl der Tupel in der Instanz
$X; X_1,...,X_n$	Zufallsvariable; Stichprobe der Größe n
n	Stichprobengröße
$Q = Q(a,b)$	Bereichsanfrage mit Bereich $[a,b]$
σ, (σ_R, σ_I)	Selektivität einer Anfrage (Relation-Selektivität, Instanz-Selektivität)
$F(x)$	Verteilungsfunktion
$f(x)$	Dichtefunktion
$F_n(x)$	empirische Verteilungsfunktion
$\hat{f}$, $\hat{\sigma}$, $\hat{h}$	Schätzung der Dichte f, bzw. der Selektivität σ, bzw. der asymptotisch optimalen Bandbreite h_{AO}
EW, ED, FP, ASH	Equi-Width (-Histogramm), Equi-Depth (-Histogramm), Häufigkeitspolygon (Frequency Polygon), Average Shifted Histogramm
(A)MISE	(approximated) mean integrated squared error
H, h, h_{OA}, h_{NS}	Bandbreiten-Matrix, Bandbreite, asymptotisch optimale Bandbreite, Bandbreite aufgrund von Normalskalierungsregel
$K(t)$	Kernfunktion
K^P, K^S, K^L, K^R	Produktkern, Sphärischer Kern, linker bzw. rechter Randkern

Tabelle A.1: Notation

A.2 Statistische Merkmale der Testdaten

Im den folgenden Tabellen sind statistische Eigenschaften der eindimensionalen bzw. zweidimensionalen (künstlichen und realen) Testdaten (Instanzen und Stichproben mit 2000 Werten) aufgeführt. Für eine Erklärung der einzelnen statistischen Begriffe sei auf die einschlägige Literatur verwiesen (z.B. [Bosch 94]). Die Werte wurden mit Microsoft-Excel'97 berechnet.

A.2.1 Univariate Testdaten

	u(20)		u(15)		n(20)	
	Instanz	**Stich-probe**	**Instanz**	**Stich-probe**	**Instanz**	**Stich-probe**
Anzahl	100.000	2.000	100.000	2.000	100.000	2.000
Wertebereich	$0..2^{20}$-1		$0..2^{15}$-1		$0..2^{20}$-1	
Minimum	28	2055	0	22	5	3922
Maximum	1.048.572	1.048.269	32.766	32.765	1.048.545	1.047.245
Mittelwert	522.855,1	522.630,8	16.361,3	16.409,8	524.765,5	524.854,8
Median	521.493,0	521.003,5	16.357,0	16.308,5	525.756,5	526.211,5
Interquartilsabstand	524.842,5	542.142,5	16.416,3	16.291,0	337.232,3	332.503,5
Standardabweichung	303.308,4	302.161,7	9.463,4	9.359,1	231.146,4	229.120,0
Schiefe	-1,202	-1,217	-1,201	-1,183	-0,637	-0,618
Kurtosis	0,008	-0,009	0,001	-0,002	-0,005	0,019

Tabelle A.2: Statistische Merkmale der univariaten Testdaten u(20), u(15) und n(20).

	n(15)		e(20)		e(15)	
	Instanz	**Stich-probe**	**Instanz**	**Stich-probe**	**Instanz**	**Stich-probe**
Anzahl	100.000	2.000	100.000	2.000	100.000	2.000
Wertebereich	$0..2^{15}$-1		$0..2^{20}$-1		$0..2^{15}$-1	
Minimum	0	44	0	49	0	1
Maximum	34.941	32.760	1.048.572	621.598	32.766	19.424
Mittelwert	16.370,4	16.151,7	87.132,7	85.664,5	2.722,3	2.676,4
Median	16.366,0	16.090,0	60.538	58.065	1.891	1.814
Interquartilsabstand	10.451,0	10.715,5	95.945,5	90.737,5	2.998	2.835,5
Standardabweichung	7.196,3	7.201,5	86.760,3	86.155,9	2.711,2	2.692,3
Schiefe	-0,629	-0,625	5,903	5,162	5,903	5,163
Kurtosis	0,003	0,037	1,980	1,975	1,980	1,975

Tabelle A.3: Statistische Merkmale der univariaten Testdaten n(15), e(20) und e(15).

	la(16)		ar1(21)		ar2(18)	
	Instanz	**Stich-probe**	**Instanz**	**Stich-probe**	**Instanz**	**Stich-probe**
Anzahl	28.136	2.000	52.120	2.000	52.120	2.000
Wertebereich	$0..2^{16}$-1		$0..2^{21}$-1		$0..2^{18}$-1	

Tabelle A.4: Statistische Merkmale der univariaten Testdaten la(16), ar1(21), und ar2(18).

	la(16)		ar1(21)		ar2(18)	
	Instanz	**Stich-probe**	**Instanz**	**Stich-probe**	**Instanz**	**Stich-probe**
Minimum	0	43	0	311	0	1.302
Maximum	65.533	65.338	2.097.151	2097057	262.143	261.938
Mittelwert	29.687,8	29.581,1	428.412,8	440.796,8	118.873,2	117.384,9
Median	30.233	29.654	318.239	321.352	111.111	110.666
Interquartilsabstand	33.598,0	33.617,5	26.889,0	260.185,3	114.862,0	115.933,0
Standardabweichung	18.772,9	18.725,5	438.837,3	450.048,1	71.684,9	72.681,7
Schiefe	-1,22	-1,23	3,49	2,91	-0,96	-0,98
Kurtosis	0,071	0,075	2,048	1,930	0,284	0,299

Tabelle A.4: Statistische Merkmale der univariaten Testdaten la(16), ar1(21), und ar2(18).

	rr1(22)		rr2(22)	
	Instanz	**Stichprobe**	**Instanz**	**Stichprobe**
Anzahl	257.942	2.000	257.942	2.000
Wertebereich	$0..2^{22}-1$		$0..2^{22}-1$	
Minimum	0	64.745	0	67.033
Maximum	4.194.303	4.039.244	4.194.303	4.194.228
Mittelwert	2.467.681,2	2.586.133,0	2.265.276,5	2.263.357,0
Median	2.479.859	2.629.043	2.316.175	2.301.454
Interquartilsabstand	1.477.649	1.437.053	1.400.763	1.433.595
Standardabweichung	924214,3	912.633,0	865102,5	878.680,1
Schiefe	-0,885	-0,429	0,173	0,167
Kurtosis	-0,263	-0,770	-0,970	-0,960

Tabelle A.5: Statistische Merkmale der univariaten Testdaten rr1(22) und rr2(22).

	rr1(12)		rr2(12)	
	Instanz	**Stichprobe**	**Instanz**	**Stichprobe**
Anzahl	257.942	2.000	257.942	2.000
Wertebereich	$0..2^{12}-1$		$0..2^{12}-1$	
Minimum	0	63	0	65
Maximum	4.095	3.943	4.095	4.094
Mittelwert	2.522,8	2.524,4	2.211,1	2.209,3
Median	2.567,0	2.566,5	2.261,0	2.246,0

Tabelle A.6: Statistische Merkmale der univariaten Testdaten rr1(12) und rr2(12).

	rr1(12)		rr2(12)	
	Instanz	**Stichprobe**	**Instanz**	**Stichprobe**
Interquartilsabstand	144.9,0	1.403,0	1.368,0	1.400,0
Standardabweichung	900,9	891,0	844,6	857,9
Schiefe	-0,417	-0,429	0,173	0,167
Kurtosis	-0,813	-0,770	-0,970	-0,960

Tabelle A.6: Statistische Merkmale der univariaten Testdaten rr1(12) und rr2(12).

A.2.2 Bivariate Testdaten

	u(15x15)				n(15x15)			
	Instanz		**Stichprobe**		**Instanz**		**Stichprobe**	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
Anzahl	100.000		2.000		100.000		2.000	
Wertebereich	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1
Minimum	0	0	19	1	0	0	139	168
Maximum	32.767	327.67	327.65	327.18	327.65	32.767	32.653	32.741
Mittelwert	16.340,5	16.390,0	16.277,1	16.309,4	16.403,9	16.394,7	16.003,2	16.392,3
Median	16.307,0	16.464,0	16.155,5	16.496,0	16.419,5	16.377,0	16.078,5	16.180,0
Interquartilsabstand	16.330,0	16.364,3	15.812,3	16.646,5	10.505,3	10.481,0	10.500,5	10.257,3
Standardabweichung	9.460,4	9.445,8	9.342,6	9.461,9	7.217,4	7.210,6	7.178,1	7.172,7
Korrelation	-0,002		0,017		0,007		0,028	

Tabelle A.7: Statistische Merkmale der bivariaten Testdaten u(15x15) und n(15x15).

	uk(15x15)				e(15x15)			
	Instanz		**Stichprobe**		**Instanz**		**Stichprobe**	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
Anzahl	100.000		2.000		100.000		2.000	
Wertebereich	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1	$0..2^{15}$-1
Minimum	0	0	0	0	0	0	0	3
Maximum	32.767	32.767	32.760	32.767	32.767	32.767	24.436	1.8656

Tabelle A.8: Statistische Merkmale der bivariaten Testdaten uk(15x15) und e(15x15).

	uk(15x15)				**e(15x15)**			
	Instanz		**Stichprobe**		**Instanz**		**Stichprobe**	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
Mittelwert	16.340,5	16.343,9	16.284,6	16.261,1	2.974,2	2.279,3	2.951,4	2,332,6
Median	16.307,0	16.301,9	16.237,0	16.282,5	2.057,0	1.581,0	2.062,5	1.659,5
Interquartilsabstand	16.330,0	16.370,8	16.534,8	16.212,3	3.281,0	2.510,0	3.175,0	2.611,8
Standardabweichung	9.460,4	9.514,6	9.478,6	9.517,3	2.964,8	2.277,7	2.897,4	2.287,3
Korrelation	0,989		0,989		-0,005		0,017	

Tabelle A.8: Statistische Merkmale der bivariaten Testdaten uk(15x15) und e(15x15).

	la(16x16)				**ar(21x18)**			
	Instanz		**Stichprobe**		**Instanz**		**Stichprobe**	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
Anzahl	28.136		2.000		52120		2.000	
Wertebereich	$0..2^{16}$-1	$0..2^{16}$-1	$0..2^{16}$-1	$0..2^{16}$-1	$0..2^{21}$-1	$0..2^{18}$-1	$0..2^{21}$-1	$0..2^{18}$-1
Minimum	0	323	24	2479	0	0	311	419
Maximum	65533	65522	65491	65482	2097151	262143	2097065	261987
Mittelwert	29687,8	34328,1	26710,9	34196,0	428412,8	118873,2	437474,0	120598,1
Median	30233,0	34707,0	25274,0	34763,0	318239,0	111111,0	270966,0	113828,0
Interquartilsabstand	33598,0	28738,0	34269,3	26160,5	260889,0	114862,0	305945,0	112410,0
Standardabweichung	18772,9	17283,5	19008,7	16718,7	438837,3	71684,9	494727,8	72102,9
Korrelation	0,247		0,236		0,195		0,284	

Tabelle A.9: Statistische Merkmale der bivariaten Testdaten la(16x16) und ar(21x18).

	rr(22x22)				**rr(12x12)**			
	Instanz		**Stichprobe**		**Instanz**		**Stichprobe**	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
Anzahl	257.942		2.000		257.942		2.000	
Wertebereich	$0..2^{22}$-1	$0..2^{22}$-1	$0..2^{22}$-1	$0..2^{22}$-1	$0..2^{12}$-1	$0..2^{12}$-1	$0..2^{12}$-1	$0..2^{12}$-1

Tabelle A.10: Statistische Merkmale der bivariaten Testdaten rr(22x22) und rr(12x12).

	rr(22x22)				rr(12x12)			
	Instanz		Stichprobe		Instanz		Stichprobe	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
Minimum	0	0	280687	617288	0	0	171	281
Maximum	4194303	4194303	4068088	4190176	4095	4095	3988	4094
Mittelwert	2584504,5	2265276,5	2930052,2	1932188,5	2522,8	2211,1	2544,8	2214,2
Median	2629302,0	2316175,0	3197527,0	1572338,5	2567,0	2261,0	2607,0	2257,5
Interquartilsabstand	1483843,8	1400763,0	1177063,0	1396876,5	1449,0	1368,0	1478,3	1381,3
Standardabweichung	922764,8	865102,5	846262,0	853174,7	900,9	844,6	894,0	851,1
Korrelation	-0,449		-0,627		-0,631		-0,431	

Tabelle A.10: Statistische Merkmale der bivariaten Testdaten rr(22x22) und rr(12x12).

A.3 Bandbreiten

Die folgenden Tabellen geben zu den Testdatensätzen die durch Experimente gefundenen optimalen (OGB) die nach einem bestimmten Verfahren geschätzten asymptotisch optimalen und Bandbreiten (AOB) wieder. Die aufgeführten Bandbreitenschätzer sind die Normalskalierungsregel (AOB-NS), die Normalskalierungsregel unter Berücksichtigung des Korrelationskoeffizienten bei bivariaten Testdaten (AOB-NS-cor) und die direkte Plug-In Regel 2. Stufe (AOB-DPI2).

	EW-HS		ASHS (s=10)		KS		
Datensatz	**OGB**	**AOB-NS**	**OGB**	**AOB**	**OGB**	**AOB-NS**	**AOB-DPI2**
u(20)	3	13	1	6	> 400.000	155.527	91.573
u(15)	1	13	1	6	> 10.000	4.853	3.193
n(20)	16	20	8	8	136.800	118.524	124.446
n(15)	15	20	9	8	3.840	3.690	3.734
e(20)	56	69	32	28	32.000	36.715	15.490
e(15)	61	69	31	28	1.040	1.147	484
la1(16)	31	13	25	6	2.304	9.626	4.962
ar1(21)	102	51	83	19	20.400	99.833	47.480
ar2(18)	94	14	54	6	2.940	36.785	22.232
rr1(22)	74	18	46	8	84.000	473.164	248.454
rr2(22)	51	19	60	8	60.000	443.597	243.296
rr1(12)	42	18	36	8	84	462	243

	EW-HS		ASHS (s=10)		KS		
Datensatz	**OGB**	**AOB-NS**	**OGB**	**AOB**	**OGB**	**AOB-NS**	**AOB-DPI2**
rr2(12)	51	18	60	8	56	424	236

Tabelle A.11: Anzahl Bins bzw. Bandbreiten im univariaten Fall.

	EW-HS						ASHS	
	OGB		AOB-NS		AOB-NS-corr		OGB	
Datensatz	**k1**	**k2**	**k1**	**k2**	**k1**	**k2**	**k1**	**k2**
u(15x15)	3	3	7	7	7	7	2	2
n(15x15)	8	8	10	10	10	10	6	6
e(15x15)	30	40	33	34	33	34	44	55
uk(15x15)	40	40	7	7	28	28	12	12
la(16x16)	13	15	7	8	7	8	14	16
ar(21x18)	65	21	27	8	27	8	27	16
rr(22x22)	42	47	9	10	10	11	18	20
rr(12x12)	47	52	9	10	11	12	18	20

Tabelle A.12: Anzahl Bins im bivariaten Fall.

	KS					
	OGB		AOB-NS		AOB-NS-corr	
Datensatz	**h1**	**h2**	**h1**	**h2**	**h1**	**h2**
u(15x15)	11.800	11.800	5.901	5.892	5.901	5.892
n(15x15)	5.100	5.100	4.502	4.498	4.502	4.497
e(15x15)	1.230	943	1.518	1.161	1.518	1.161
uk(15x15)	795	795	5.901	5.935	1.345	1.353
la(16x16)	4.410	4.074	11.709	10.780	11.465	10.555
ar(21x18)	28.522	10.478	120.718	44.713	119.171	44.140
rr(22x22)	46.791	43.866	575.568	539.602	533.464	500.129
rr(12x12)	55	50	562	527	472	442

Tabelle A.13: Bandbreiten im bivariaten Fall.

A.4 Testergebnisse

Die folgenden Testergebnisse beziehen sich - wenn nicht anders gesagt - auf eine Stichprobe der Größe n = 2000 und Anfragegrößen von 1% der Domänengröße.

A.4.3 Univariate Testergebnisse

Testergebnisse bei Gleichverteilungsannahme und direkter Selektivitätsschätzung

In Tabelle A.14 ist der mittlere relative Selektivitätsfehler für die Selektivitätsschätzung mit Gleichverteilungsannahme und für den direkten Selektivitätsschätzer aufgeführt.

Testdatensatz	GS	DS
u(20)	2,89%	17,9%
u(15)	2,39%	18,0%
n(20)	37,0%	15,8%
n(15)	41,7%	17,6%
e(20)	104,4%	9,7%
e(15)	206,9%	10,3%
la1(16)	19,9%	14,8%
ar1(21)	126,4%	15,2%
ar2(18)	23,1%	18,5%
rr1(22)	45,8%	14,7%
rr2(22)	50,3%	14,5%
rr1(12)	43,2%	14,9%
rr2(12)	44,7%	14,2%

Tabelle A.14: MRSF von GS und DS.

Testergebnisse der verschiedenen Histogrammselektivitätsschätzer

In Tabelle A.15 ist der mittlere relative Selektivitätsfehler für alle drei Histogrammselektivitätsschätzer mit optimaler gefundener Anzahl Bins aufgeführt.

Testdatensatz	EW-HS(OGB)	ED-HS(OGB)	MD-HS(OGB)
u(20)	2,87%	2,88%	2,89%
u(15)	2,39%	2,39%	2,37%
n(20)	5,70%	6,77%	8,32%
n(15)	6,44%	9,49%	8,90%
e(20)	5,85%	9,14%	8,94%

Tabelle A.15: MRSF des HS und verschiedenen Histogrammtypen mit jeweils OGB.

Testdatensatz	EW-HS(OGB)	ED-HS(OGB)	MD-HS(OGB)
e(15)	5,96%	8,14%	8,75%
la1(16)	8,00%	7,79%	8,94%
ar1(21)	11,54%	12,08%	13,78%
ar2(18)	16,12%	15,01%	17,15%
rr1(22)	10,57%	13,71%	11,71%
rr2(22)	10,32%	12,25%	12,05%
rr1(12)	10,53%	13,13%	11,52%
rr2(12)	10,67%	12,16%	12,45%

Tabelle A.15: MRSF des HS und verschiedenen Histogrammtypen mit jeweils OGB.

In Tabelle A.16 ist der mittlere relative Selektivitätsfehler für alle drei Histogrammselektivitätsschätzer aufgeführt. Die Anzahl der Bins ist bei allen drei Schätzern gleich und entspricht der bei dem EW-HS mit Normalskalierungsregel geschätzter asymptotisch optimaler Anzahl Bins.

Testdatensatz	EW-HS(AOB)	ED-HS	MD-HS
u(20)	8,5%	7,9%	6,6%
u(15)	5,6%	7,5%	7,4%
n(20)	6,8%	7,2%	15,2%
n(15)	8,4%	9,7%	23,0%
e(20)	7,3%	8,9%	40,5%
e(15)	7,7%	11,5%	39,6%
la1(16)	9,0%	9,9%	10,9%
ar1(21)	14,4%	12,3%	29,9%
ar2(18)	17,6%	17,7%	18,8%
rr1(22)	13,0%	19,6%	22,6%
rr2(22)	17,1%	20,7%	21,8%
rr1(12)	12,9%	15,7%	21,8%
rr2(12)	16,7%	18,5%	23,3%

Tabelle A.16: MRSF des HS bei verschiedenen Histogrammtypen mit AOB-NS für EW-HS.

In Tabelle A.17 ist der mittlere relative Selektivitätsfehler für den Average Shifted Histogrammselektivitätsschätzer mit optimaler gefundener und mit geschätzter asymptotisch optimaler Anzahl Bins sowie mit jeweils 10 Shifts aufgeführt.

Testdatensatz	ASHS(OGB,10)	ASHS(AOB,10)
u(20)	2,87%	6,2%
u(15)	2,46%	6,0%
n(20)	3,13%	4,8%
n(15)	3,13%	7,0%
e(20)	4,86%	6,1%
e(15)	4,82%	6,5%
la1(16)	7,67%	8,03%
ar1(21)	11,69%	13,5%
ar2(18)	15,56%	16,5%
rr1(22)	10,78%	13,8%
rr2(22)	10,73%	13,7%
rr1(12)	10,30%	11,9%
rr2(12)	10,94%	13,5%

Tabelle A.17: MRSF des ASHS mit OGB und AOB.

Testergebnisse der verschiedenen Kernselektivitätsschätzer

In Tabelle A.18 ist der mittlere relative Selektivitätsfehler für alle drei Kernselektivitätsschätzer (ohne Randbehandlung, mit Spiegelung, mit speziellem Randkern) bei optimaler gefundener Bandbreite aufgeführt.

Testdatensatz	KS-O(OGB)	KS-S(OGB)	KS-R(OGB)
u(20)	8,3%	2,85%	3,0%
u(15)	7,9%	2,8%	3,3%
n(20)	3,06%	3,09%	3,01%
n(15)	4,65%	4,9%	4,57%
e(20)	7,2%	4,9%	4,68%
e(15)	7,6%	4,67%	4,45%
la1(16)	8,5%	7,96%	8,05%
ar1(21)	11,75%	11,82%	11,79%
ar2(18)	15,9%	15,9%	15,9%
rr1(22)	11,04%	11,04%	11,05%

Tabelle A.18: MRSF des KS mit und ohne Randbehandlung mit jeweils OGB.

Testdatensatz	KS-O(OGB)	KS-S(OGB)	KS-R(OGB)
rr2(22)	10,76%	10,76%	10,76%
rr1(12)	10,64%	10,64%	10,64%
rr2(12)	11,50%	11,50%	11,50%

Tabelle A.18: MRSF des KS mit und ohne Randbehandlung mit jeweils OGB.

Tabelle A.19 ist der mittlere relative Selektivitätsfehler für alle drei Kernselektivitätsschätzer (ohne Randbehandlung, mit Spiegelung, mit speziellem Randkern) bei mit Normalskalierungsregel geschätzter asymptotisch optimaler Bandbreite aufgeführt.

Testdatensatz	KS-O(AOB)	KS-S(AOB)	KS-R(AOB)
u(20)	8,4%	4,0%	4,1%
u(15)	8,7%	4,2%	4,7%
n(20)	3,16%	3,14%	3,22%
n(15)	4,67%	4,9%	4,59%
e(20)	9,5%	5,0%	4,76%
e(15)	9,5%	4,71%	4,46%
la1(16)	14,5%	9,6%	9,1%
ar1(21)	22,9%	23,6%	22,7%
ar2(18)	17,9%	17,2%	17,0%
rr1(22)	21,1%	21,6%	18,9%
rr2(22)	19,9%	20,1%	19,9%
rr1(12)	17,8%	17,8%	17,6%
rr2(12)	20,3%	20,4%	20,3%

Tabelle A.19: MRSF des KS mit und ohne Randbehandlung mit jeweils AOB.

Testergebnisse der verschiedenen Verfahren zur Bandbreitenbestimmung

In Tabelle A.20 ist der mittlere relative Selektivitätsfehler für die verschiedenen Kernselektivitätsschätzer mit und ohne Randbehandlung bei einer durch die Direkte Plug-In Regel geschätzten asymptotisch optimalen Bandbreite sowie der Kernselektivitätsschätzer ohne Randbehandlung bei einer adaptiven Bandbreite aufgeführt.

Testdatensatz	KS-O(DPI2)	KS-S(DPI2)	KS-R(DPI2)	KS-O(AD)
u(20)	8,28%	5,54%	5,75%	10,46%
u(15)	7,95%	4,96%	5,43%	9,54%
n(20)	3,11%	3,10%	3,13%	4,42%
n(15)	4,67%	4,90%	4,58%	5,51%

Tabelle A.20: MRSF des KS mit erweiterten Verfahren zur Bestimmung der Bandbreite.

Testdatensatz	KS-O(DPI2)	KS-S(DPI2)	KS-R(DPI2)	KS-O(AD)
e(20)	7,32%	5,95%	5,86%	6,82%
e(15)	7,58%	6,15%	6,02%	6,50%
la1(16)	10,78%	8,46%	8,56%	11,32%
ar1(21)	15,03%	15,49%	15,18%	13,16%
ar2(18)	17,25%	16,91%	16,96%	17,90%
rr1(22)	15,08%	15,09%	14,41%	15,44%
rr2(22)	15,84%	15,91%	15,88%	15,72%
rr1(12)	13,06%	13,06%	13,05%	13,49%
rr2(12)	16,12%	16,18%	16,14%	15,67%

Tabelle A.20: MRSF des KS mit erweiterten Verfahren zur Bestimmung der Bandbreite.

Testergebnisse des Hybridselektivitätsschätzers

Tabelle A.21 zeigt den mittleren relativen Selektivitätsfehler für den Hybridselektivitätsschätzer mit Spiegelung und speziellem Randkern innerhalb der Histogrammbins sowie zum Vergleich des Histogrammselektivitätsschätzers bei jeweils heuristisch bestimmter Bingrenzen.

Testdatensatz	*k*	YS-S(x)	YS-R(x)	HS(x)
xu(15)	9	7,8%	8,7%	3,1%
xe(15)	9	7,0%	6,8%	84,8%
la1(16)	2	8,6%	8,6%	20,6%
ar1(21)	11	13,1%	13,0%	28,3%
ar2(18)	12	16,0%	14,9%	17,2%
rr1(22)	11	12,1%	12,2%	19,1%
rr2(22)	8	11,9%	11,1%	27,5%
rr1(12)	11	11,7%	11,4%	17,8%
rr2(12)	9	12,2%	11,0%	22,6%

Tabelle A.21: MRSF des Hybridselektivitätsschätzers und korrespondierenden HS.

A.4.4 Bivariate Testergebnisse

Testergebnisse bei Gleichverteilungsannahme und direkter Selektivitätsschätzung

In Tabelle A.22 ist der mittlere relative Selektivitätsfehler für die Selektivitätsschätzung mit Gleichverteilungsannahme und für den direkten Selektivitätsschätzer aufgeführt.

Testdatensatz	GS	DS
u(15x15)	2,47%	17,98%
n(15x15)	48,09%	15,39%
e(15x15)	129,14%	8,78%
uk(15x15)	79,97%	7,69%
la(16x16)	26,72%	15,41%
ar(21x18)	136,93%	11,71%
rr(22x22)	67,88%	12,42%
rr(12x12)	66,78%	8,74%

Tabelle A.22: MRSF von GS und DS.

Testergebnisse der verschiedenen Histogrammselektivitätsschätzer

In Tabelle A.23 ist der mittlere relative Selektivitätsfehler des Equi-Width- und des Equi-Depth-Histogrammselektivitätsschätzers aufgeführt. Die Bandbreite wurde beim Equi-Width-Histogrammselektivitätsschätzer nach der Normalskalierungsregel geschätzt und ebenfalls beim Equi-Depth-Histogrammselektivitätsschätzer angewendet.

Testdatensatz	EW-HS(AOB)	ED-HS
u(15x15)	7,38%	9,53%
n(15x15)	9,34%	13,41%
e(15x15)	7,54%	11,69%
uk(15x15)	28,59%	32,63%
la(16x16)	16,34%	20,11%
ar(21x18)	16,17%	15,66%
rr(22x22)	27,62%	29,19%
rr(12x12)	24,79%	26,91%

Tabelle A.23: MRSF von EW-HS(AOB) und ED-HS.

In Tabelle A.24 ist der mittlere relative Selektivitätsfehler des Equi-Width- und des Equi-Depth-Histogrammselektivitätsschätzers bei OGB aufgeführt.

Testdatensatz	EW-HS(OGB)	ED-HS(OGB)
u(15x15)	4,09%	4,01%
n(15x15)	8,55%	12,98%
e(15x15)	7,19%	8,89%
uk(15x15)	6,77%	7,19%
la(16x16)	13,67%	13,68%
ar(21x18)	11,15%	11,76%
rr(22x22)	12,45%	13,31%
rr(12x12)	8,38%	10,83%

Tabelle A.24: MRSF von EW-HS(OGB) und ED-HS(OGB).

In Tabelle A.25 ist der mittlere relative Selektivitätsfehler des Average Shifted Histogrammselektivitätsschätzers bei OGB aufgeführt.

Testdatensatz	ASHS(OGB)
u(15x15)	3,0%
n(15x15)	6,5%
e(15x15)	7,7%
uk(15x15)	14,9%
la(16x16)	13,5%
ar(21x18)	13,3%
rr(22x22)	14,7%
rr(12x12)	11,9%

Tabelle A.25: MRSF von ASHS(OGB) bei 2 Shifts je Dimension.

Tabelle A.26 ist der mittlere relative Selektivitätsfehler für verschiedenen Histogrammselektivitätsschätzer auf einer Z-Ordnung bei mit Normalskalierungsregel geschätzter asymptotisch optimaler Bandbreite aufgeführt.

Testdatensatz	EW-ZHS(AOB)	ED-ZHS(AOB)	MD-ZHS(AOB)
u(15x15)	5,3%	6,2%	4,7%
n(15x15)	32,2%	30,4%	22,5%
e(15x15)	19,2%	7,7%	7,8%
uk(15x15)	63,6%	70,5%	56,2%

Tabelle A.26: MRSF der verschiedenen ZHS.

Testdatensatz	EW-ZHS(AOB)	ED-ZHS(AOB)	MD-ZHS(AOB)
la(16x16)	23,7%	24,9%	30,6%
ar(21x18)	38,0%	28,4%	30,5%
rr(12x12)	51,6%	43,9%	54,4%

Tabelle A.26: MRSF der verschiedenen ZHS.

In Tabelle A.27 ist der mittlere relative Selektivitätsfehler des KD-Baum Selektivitätsschätzers bei optimal gefundener Anzahl Bins aufgeführt.

Testdatensatz	KD-HS(OGB)
u(15x15)	2,44%
n(15x15)	9,9%
e(15x15)	7,4%
uk(15x15)	7,2%
la(16x16)	14,6%
ar(21x18)	12,5%
rr(22x22)	15,2%
rr(12x12)	10,9%

Tabelle A.27: MRSF des KD-HS bei OGB.

Testergebnisse der verschiedenen Kernselektivitätsschätzer

Tabelle A.28 ist der mittlere relative Selektivitätsfehler für alle drei Kernselektivitätsschätzer (ohne Randbehandlung, mit Spiegelung, mit speziellem Randkern) bei mit Normalskalierungsregel geschätzter asymptotisch optimaler Bandbreite aufgeführt.

Testdatensatz	KS-O(AOB)	KS-S(AOB)	KS-R(AOB)
u(15x15)	9,8%	4,9%	5,3%
n(15x15)	5,8%	5,7%	5,9%
e(15x15)	7,5%	6,9%	7,2%
uk(15x15)	44,7%	42,3%	39,4%
la(16x16)	20,3%	17,7%	16,9%
ar(21x18)	21,3%	19,9%	19,2%
rr(22x22)	33,9%	33,7%	32,9%
rr(12x12)	30,8%	30,6%	29,9%

Tabelle A.28: MRSF des KS mit und ohne Randbehandlung mit jeweils AOB.

In Tabelle A.29 ist der mittlere relative Selektivitätsfehler für alle drei Kernselektivitätsschätzer (ohne Randbehandlung, mit Spiegelung, mit speziellem Randkern) bei optimaler gefundener Bandbreite aufgeführt.

Testdatensatz	KS-O(OGB)	KS-S(OGB)	KS-R(OGB)
u(15x15)	8,96%	4,03%	3,45%
n(15x15)	5,45%	5,41%	5,60%
e(15x15)	7,27%	6,67%	6,87%
uk(15x15)	6,86%	6,83%	6,86%
la(16x16)	13,03%	12,98%	12,97%
ar(21x18)	10,56%	10,63%	10,58%
rr(22x22)	12,21%	12,21%	12,21%
rr(12x12)	8,29%	8,29%	8,29%

Tabelle A.29: MRSF des KS mit und ohne Randbehandlung mit jeweils OGB.

Testergebnisse bei AOB mit Korrelationskoeffizient

In Tabelle A.30 ist der mittlere relative Selektivitätsfehler des Equi-Width-Histogramm- und der verschiedenen Kernselektivitätsschätzer bei mittels Normalskalierungsregel unter Berücksichtigung des Korrelationsfaktors geschätzter AOB aufgeführt.

Testdatensatz	EW-HS(NS-corr)	KS-O(NS-corr)	KS-S(NS-corr)	KS-R(NS-corr)
u(15x15)	7,38%	9,78%	4,87%	5,28%
n(15x15)	9,34%	5,79%	5,70%	5,87%
e(15x15)	7,54%	7,54%	6,91%	7,24%
uk(15x15)	7,40%	7,79%	7,64%	7,72%
la(16x16)	16,34%	20,32%	17,66%	16,88%
ar(21x18)	16,17%	21,25%	19,89%	19,22%
rr(22x22)	23,97%	30,82%	30,69%	30,25%
rr(12x12)	19,59%	25,00%	25,00%	24,70%

Tabelle A.30: MRSF von EW-HS und KS bei jeweils AOB-NS-corr.

A.5 Scatterplots der realen zweidimensionalen Testdaten

Scatterplot der LA-Testdaten

Das folgende Scatterplot besteht aus allen 28.136 Punkten des zweidimensionalen Los Angeles-Testdatensatzes:

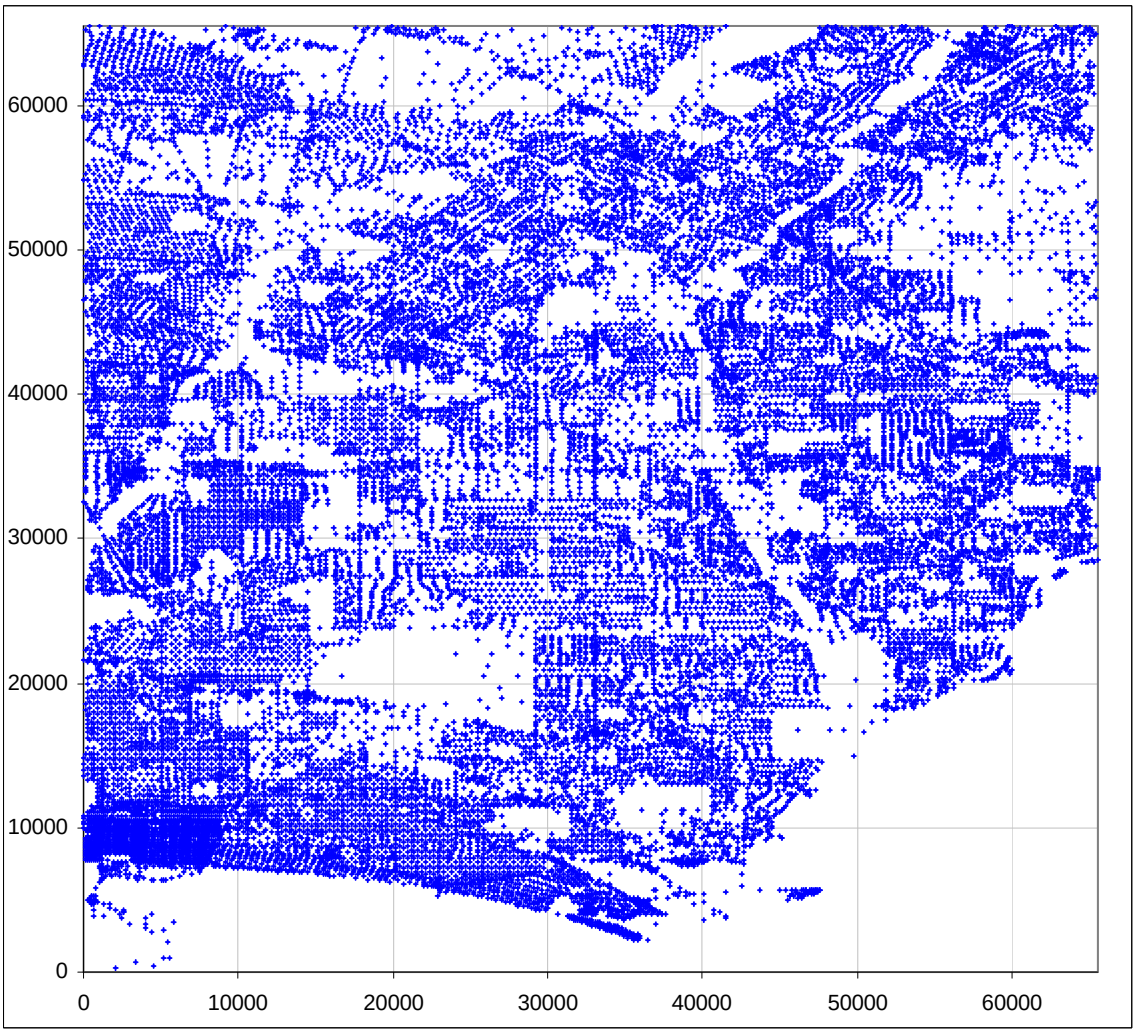


Abbildung A.1: Scatterplot der la(16x16)-Daten der Grundgesamtheit mit N = 28.136.

Scatterplot der Arapahoe-Testdaten

Das folgende Scatterplot besteht aus einer repräsentativen Stichprobe der Größe n = 30.000 des zweidimensionalen Arapahoe-Testdatensatzes:

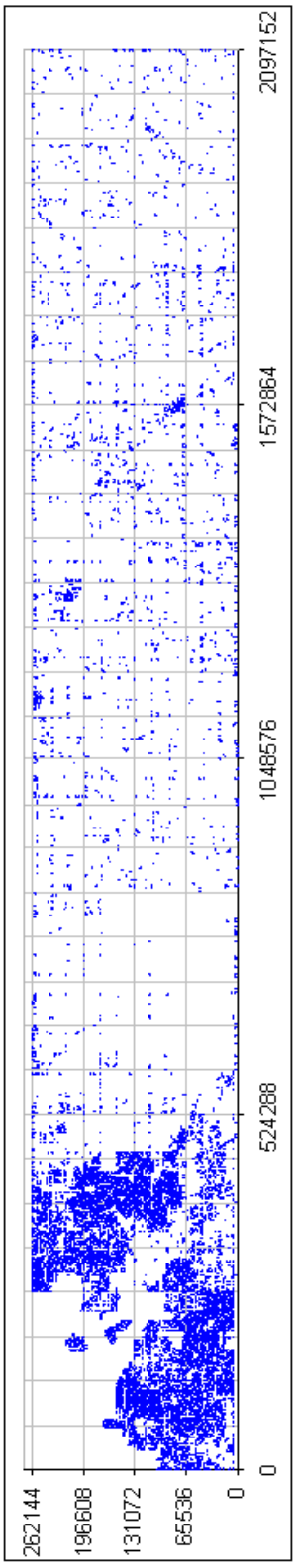


Abbildung A.2: Scatterplot der ar(21x18)-Daten bei einer Stichprobe der Größe n = 30.000.

Danksagung

Mein großer Dank gebührt meiner Frau Heike und meinen Kindern Jan-Dominik und Frederik, die mich im Laufe dieser Arbeit und im Vorfeld in mancherlei Hinsicht entbehren mußten und mir dennoch die nötige Kraft für diese Arbeit gegeben haben. Des weiteren danke ich dem Fachbereich Mathematik und Informatik und ganz besonders meinem Betreuer Bernhard Seeger, ohne deren Unterstützung diese Arbeit nicht möglich gewesen wäre. Volker Mammitzsch danke ich für seine Hinweise insbesondere im Bereich der Kernschätzer und Hans-Georg Müller (University of California at Davis) für seine Anregungen im Bereich der Randkernschätzer. Björn Blohsfeld danke ich für zahlreiche fruchtbare Diskussionen und wünsche ihm viel Erfolg bei seiner weiteren Forschungsarbeit auf diesem Themengebiet.

Quellennachweis

Schrifttum

[Aurenhammer 91] Aurenhammer, Franz „Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure" *ACM Computing Surveys* 23 (3), Sept. 1991, S. 344-405.

[Barbará et al. 97] Barbará, D. & DuMouchel, W. & Faloutsos, C. & Haas, P.J. & Hellerstein, J.M. & Ioannidis, Y. & Jagadish, H.V. & Johnson, T. & Ng, R. & Poosala, V. & Ross, K.A. & Sevcik, K.C. „The New Jersey Data Reduction Report" Bulletin of the Technical Committee on Data Engineering, Vol. 20, No. 4, IEEE Computer Society, Dezember 1997, S. 3-45.

[Bauer 91] Bauer, H. *Wahrscheinlichkeitstheorie* Walter de Gruyter, 1991.

[Beckmann et al. 90] Beckmann, N. & Kriegel, H.P. & Schneider, R, & Seeger, B. „The R*-tree: An efficient and robust access method for points and rectangles" *ACM SIGMOD*, Atlantic City, NJ, 1990.

[Belussi & Faloutsos 95] Belussi, Alberto & Faloutsos, Christos. „Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension" Proc. 21st Intl. Conf. on VLDB, Zürich 1995, S. 299-310.

[Bosch 89] Bosch, Karl *Elementare Einführung in die Wahrscheinlichkeitsrechnung* vieweg 1986[5].

[Bosch 94] Bosch, Karl *Elementare Einführung in die angewandte Statistik* vieweg 1994[5].

[Blohsfeld 98] Blohsfeld, Björn, *Selektivitätsschätzung von mehrdimensionalen Datenbankanfragen mit Kernschätzern*, Diplomarbeit in Wirtschaftsmathematik, Philipps-Universität Marburg, Mai 1998.

[Blohsfeld et al. 99] Blohsfeld, Björn & Korus, Dieter & Seeger, Bernhard "A Comparison of Selectivity Estimators for Range Queries on Metric Attributes" Proc. SIGMOD'99, Philadelphia 1999.

[Büning & Trenkler 94] Büning, H. & Trenkler, G. *Nichtparametrische statistische Methoden* Walter de Gruyter 1994.

[Buskamp 97] Buskamp, Herbert, *Histogramme zur Schätzung der Selektivität mehrdimensionaler Anfragen*, Diplomarbeit in Wirtschaftsinformatik, Philipps-Universität Marburg, Februar 1997.

[Chen & Roussopoulos 94] Chen, Chungmin Melvin & Roussopoulos, Nick „Adaptive Selectivity Estimation Using Query Feedback" *Proc. ACM-SIGMOD Intl. Conf. on Management of Data*, Mai 1994.

[Chiueh 94] Chiueh, T. „Content-based image-indexing" *Proc. 20th VLDB Intl. Conf.*, Santiago, Chile, Sept. 1994, S. 582-593.

[Dong & Simonoff 94] Dong, J. & Simonoff, J. „The Construction and Properties of Boundary Kernels for Sparse Multinomials" *Journal of Computational and Graphical Statistics*, Vol. 3, No. 1, 1994, S. 57-66.

[Faloutsos et al. 94] Faloutsos, Ch. & Equitz, M. & Flickner, M. & Niblack, W. & Petkovic, D. & Barber, R. „Efficient and effective querying by image content." J. of Intelligent Information Systems, 3 (3/4), Juli 1994, S. 231-262.

[Faloutsos & Kamel 94] Faloutsos, Christos & Kamel, Ibrahim „Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension" Proc. ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems PODS, Minneapolis, MN, Mai 1994, S. 4-13.

[Faloutsos & Roseman 89] Faloutsos, Christos & Roseman, Shari „Fractals for Secondary Key Retrieval" *Proc. 8th Symp. on Principals of Database Systems PODS'89*, 1989, S. 247-252.

[Fayyad et al. 96] Fayyad, Usama & Piatetsky-Shapiro, Gregory & Smyth, Padhraic „The KDD Process for Extracting Useful Knowledge from Volumens of Data" *Communications of the ACM* Vol. 39, No. 11, Nov. 1996, S. 27-34.

[Forster 79] Forster, Otto *Analysis* 1, vieweg 1979.

[Gasser et al. 93] Gasser, T. & Engel, J. & Seifert, B. „Nonparametric function estimation" in: Rao (Ed.), *Handbook of Statistics* Vol. 9, Kap. 12, North Holland 1993.

[Gasser & Müller 79] Gasser, T. & Müller, H.-G. „Kernel estimation of regression functions" in T. Gasser & M. Rosenblatt (Hrsg.), Springer 1979, S. 23-68.

[Gibbons & Matias 98] Gibbons, Phillip B. & Matias, Yossi „New Sampling-Based Summary Statistics for Improving Approximate Query Answers" SIGMOD Conf. 1998, S. 331-342.

[Glymour et al. 96] Glymour, Clark & Madigan, David & Pregibon, Daryl & Smyth, Padhraic „Statistical Inference and Data Mining" *Communications of the ACM* Vol. 39, No. 11, Nov. 1996, S. 35-41.

[Granovsky et al. 95] Granovsky, B.L. & Müller, H.-G. & Pfeiffer, C. „Some Remarks on Optimal Kernel Functions" *Statistics & Decisions* 13, 1995, S. 101-116.

[Güting 94] Güting, Ralf Hartmut „An introduction to spatial database systems" *VLDB Journal* 3, 1994, S. 357-399.

[Guttmann 84] Guttmann, A. „R-trees: A dynamic index structure for spatial searching" *Proc. of the ACM SIGMOD*, 1984.

[Haas & Swami 92] Haas, P. & Swami, A. „Sequential Sampling Procedures for Query Size Estimation" Proc. ACM SIGMOD Intl. Conf. on Management of Data, San Diego, CA, 1992, S. 341-350.

[Hartigan 75] Hartigan, J.A., *Clustering Algorithms*, Wiley, New York 1975.

[Hartung & Elpert 95] Hartung, Joachim & Elpert, Bärbel *Multivariate Statistik*, Oldenbourg Verlag 1995.

[Hellerstein et al. 97] Hellerstein, Joseph M. & Haas, Peter J. & Wang, Helen: „Online Aggregation" SIGMOD Conference 1997, S. 171-182.

[Hellerstein & Pfeffer 94] Hellerstein, J. M. & Pfeffer, A. „The RD-tree: An index structure for sets" Tech. Rep. 1252, Univ. of Wisconsin at Madison, Oktober 1994.

[Holsheimer & Kersten 94] Holsheimer, Marcel & Kersten, Martin L. „Architectural Support for Data Mining" Tech. Rep. CS-R9429, Centrum voor Wiskunde en Informatica, Amsterdam 1994.

[Imielinski & Mannila 96] Imielinski, Tomasz & Mannila, Heikki „A Database Perspective on Knowledge Discovery" *Communications of the ACM* Vol. 39, No. 11, Nov. 1996, S. 58-64.

[Ioannidis & Christodoulakis 91] Ioannidis, Y.E. & Christodoulakis, S. „On the Propagation of Errors in the Size of Join Results" *Proc ACM-SIGMOD Intl. Conf. on Management of Data*, Denver 1991, S. 268-277.

[Ioannidis & Poosala 95] Ioannidis, Yannis E. & Poosala, Viswanath „Balancing Histogram Optimality and Practicality for Query Result Size Estimation" Proc. ACM-SIGMOD Conf. Mai 1995, S. 233-244.

[Jones 93] Jones, M.C. „Simple Boundary Correction for Kernel Density Estimation" *Statistics and Computing* 3, 1993, S. 135-146.

[Jones & Foster 96] Jones, M.C. & Foster, P.J. „A Simple Nonnegative Boundary Correction Method for Kernel Density Estimation" *Statistica Sinica*, Vol. 6, No. 4, Oktober 1996, S. 1005-1013.

[Kemper & Eickler 97] Kemper, Alfons & Eickler, André *Datenbanksysteme* 2. Aufl. Oldenbourg Verlag 1997.

[Knuth 69] Knuth, Donald E. *The Art of Computer Programming, Vol. 2 Seminumerical Algorithms*, Addison-Wesley Publ. 1969.

[Lipton & Naughton 90] Lipton, R.J. & Naughton, J.F. „Practical Selectivity Estimation Through Adaptive Sampling" Proc. ACM SIGMOD Intl. Conf. on Management of Data, Atlantic City, NJ, 1990, S. 1-11.

[Mannino et al. 88] Mannino, M.V. & Chu, Paicheng & Sager, T. „Statistical Profile Estimation in Database Systems" *ACM Computing Surveys* 20 (3) Sept. 1988, S. 191-221.

[Matias et al. 98] Matias, Y. & Vitter, J.S. & Wang, M. „Wavelet-Based Histograms for Selectivity Estimation" *Proc. ACM SIGMOD Intl. Conf. on Management of Data* (SIGMOD '98), Seattle, Washington, Juni 1998.

[Müller 88] Müller, H.-G., *Nonparametric Regression Analysis of Longitudinal Data*, Springer, Lect. Notes in Statistics 46, 1988.

[Muralikrishna & DeWitt 88] Muralikrishna, M. & DeWitt, David J. „Equi-Depth Histograms For Estimating Selectivity Factors For Multi-Dimensional Queries" *Proc. SIGMOD Intl. Conf. on Management of Data*, Chicago, Illinois, 1.-3. Juni 1988, S. 28-36.

[NCR 99] NCR Corporation „Evolving the Data Mining Process for Teradata Warehouses" White Paper, NCR Corporation, 1998-1999.

[Olkem & Rotem 94] Olken, Frank & Rotem, Doron „Random Sampling from Databases - A Survey" *Manuskript*, Information and Computing Science Div., Lawrence Berkeley Laboratory, Berkeley, CA, 22. März 1994.

[Oracle 99] Oracle SQL- und Zugriffsoptimierung Schulungsunterlagen K1110 V4.0, Oracle Deutschland GmbH, März 1999.

[Ottmann & Widmayer 96] Ottmann, T. & Widmayer, P. „*Algorithmen und Datenstrukturen*" Spektrum, 3. Aufl. 1996.

[Pagel et al. 93] Pagel, Bernd-Uwe & Six, Hans-Werner & Toben, Heinrich & Widmayer, Peter „Towards an Analysis of Range Query Performance in Spatial Data Structures" *Proc. ACM Principles of Database Systems*, PODS'93, Washington, 1993, S. 214-221.

[Poosala 97] Poosala, Viswanath „Histogram-based Estimation Techniques in Databases" Ph.D. Thesis, Univ. of Wisconsin-Madison 1997.

[Poosala et al. 96] Poosala, Viswanath & Ioannidis, Yannis E. & Haas, Peter J. & Shekita, Eugene J. „Improved Histograms for Selectivity Estimation of Range Predicates" *SIGMOD*'96, Montreal/Canada, 1996, S. 294-305.

[Poosala & Ioannidis 97] Poosala, Viswanath & Ioannidis, Yannis E. „Selectivity Estimation Without the Attribute Value Assumption" *Proc. 23rd VLDB Conf.*, Athen, Griechenland, 1997.

[Qiu 97] Qiu, Peihua „Nonparametric Estimation of Jump Surface" *Sankhya: The Indian Journal of Statistics* Vol. 59, Ser. A, Pt. 2, 1997.

[Qiu & Yandell 94] Qiu, Peihua & Yandell, Brian „Jump Detection in Regression Surfaces" Tech. Rep. #948, Univ. of Wisconsin, Dep. of Statistics, May 1994.

[Rao 93] Rao (Ed.), *Handbook of Statistics* Vol. 9, North Holland 1993.

[Rasch 95] Rasch, O. *Mathematische Statistik*, Verlag Johann Ambrosius Barth, 1995.

[Rojas 93] Rojas, Raúl *Theorie der neuronalen Netze* Springer 1993

[Robinson 81] Robinson, J.T. „The KDB-tree: A search structure for large multidimensional dynamic indexes" *Proc. of the ACM SIGMOD Conf.*, Ann Arbor, MI, 1981.

[Sagan 94] Sagan, Hans, *Space-Filling Curves*, Springer-Verlag 1994.

[Samet 90] Samet, H., *The Design and Analysis of Spatial Data Structures*, Reading, MA: Addison-Wesley 1990.

[Sarle 94] Sarle, Warren S. „Neural Networks and Statistical Models" *Proc. 19th Annual SAS Users Group Intl. Conf.*, April 1994.

[Schneider 97] Schneider, Carmen, *Entwicklung von Methoden zur Abschätzung der Größe von Anfragen in Geo-Datenbanken*, Diplomarbeit, Philipps-Universität Marburg, Juli 1997.

[Schreiber 91] Schreiber, Thomas „A Voronoi Diagram Based Adaptive k-Means-Type-Algorithm for Multidimensional Weighted Data" *Interner Bericht*, Fachbereich Informatik, Universität Kaiserslautern 1991.

[Scott 92] Scott, David W., *Multivariate Density Estimation*, Wiley & Sons 1992.

[Seeger 96] Seeger, Bernhard, *Index- und Speicherstrukturen*, Vorlesungsskript, Philipps-Universität Marburg, FB Mathematik, FG Informatik, SS 1996.

[Selinger et al. 79] Selinger, P.G. & Astrahan, M.M. & Chamberlin, D.D. & Lorie, R.A. & Price, T.T. „Access path selection in a relational database management system" *Proc. ACM SGMOD Conf.* 1979, S. 23-34.

[Shamos & Hoey 75] Shamos, M.I. & Hoey, D. „Closest-Point Problems" *Proc. IEEE Symp. on Foundations of Computer Science* 1975, S.151-162.

[Silverman 86] Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman & Hall 1986.

[Theodoridis & Sellis 96] Theodoridis, Yannis & Sellis, Timos „A Model for the Prediction of R-tree Performance" *Proc. 15th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems*, PODS'96, Montreal, Juni 1996, S. 161-171.

[Tropf & Herzog 81] Tropf, H. & Herzog, H. „Multidimensional Range Search in Dynamically Balanced Trees" *Angewandte Informatik* 1981.

[Vitter 85] Vitter, J.S. „Random sampling with a reservoir" *ACM Trans. Math. Software* 11, 1985, S. 37-57.

[Walter 98] Walter, Todd „NCR Teradata Version 2 Release 3 - Extending the Lead in Data Warehousing" NCR Corporation 1998.

[Wand & Jones 95] Wand, M.P. & Jones, M.C., *Kernel Smoothing*, Chapman & Hall 1995.

[Wand et al. 91] Wand, M.P. & Marron, J.S. & Ruppert, D. „Transformation in Density Estimation" *Journal of the American Statistical Association* 86, 1991, S. 343-361.

[Whang et al. 94] Whang, Kyo-Young & Kim, Sang-Wook & Wiederhold, G. „Dynamic Maintenance of Data Distribution for Selectivity Estimation" *VLDB Journal* 3, 1994, S. 29-51.

[Zezula et al. 96] Zezula, Pavel & Ciaccia, Paolo & Rabitti, Fausto "M-tree: A Dynamic Index for Similarity Queries in Multimedia Databases" *Tech. Rep*. 7, HERMES ESPRIT LTR Proj. 1996.

[Zipf 49] Zipf, G.K. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*, Addison Wesley 1949.

Internetadressen

Man beachte, daß Internetadressen einem steten Wandel unterzogen sind. Eine Adresse, die heute noch gültig ist, kann morgen schon nicht mehr existieren. Die unten angegebenen Adressen sind daher ohne Gewähr.

[census data] http://www.census.gov/

[teradata] http://www.ncr.com/

[terra server] http://terraserver.microsoft.com

[tiger data] http://www.census.gov/ftp/pub/geo/www/tiger

Liste der Abbildungen

Kapitel 1

Kapitel 2

Kapitel 3

Kapitel 6

Kapitel A

Liste der Tabellen

Kapitel 1

Kapitel 2

Kapitel 3

Kapitel 4

Kapitel 5

Kapitel 6

Kapitel A

Index

Lebenslauf

Dieter Korus

29.05.1961	geboren in Brilon
	verheiratet, zwei Söhne
1967 - 1971	Grundschule Goetheschule Herten i.Westf.
1971 - 1980	Mathematisch-Naturwissenschaftliches Gymnasium Herten, Abitur
08/1980 - 02/1984	Studium Lehramt Sekundarstufe II, Mathematik und Philosophie an der Wilhelms-Universität Münster i.Westf., Zwischenexamen
02/1984 - 09/1986	Studium Lehramt Gymnasium, Mathematik und Sozialkunde an der Philipps-Universität Marburg, Zwischenexamen
09/1986 - 07/1991	Studium Diplom-Mathematik mit Nebenfach Informatik an der Philipps-Universität Marburg
07/1991	Diplom in Mathematik, Diplomarbeit im Nebenfach Informatik über "Näherungslösungen des Travelling Salesman Problems mit Neuronalen Netzen"
05/1992 - 12/1992	wissenschaftlicher Angestellter im BMBF-Projekt "WiNA - Wissensverarbeitung in neuronaler Architektur" an der Universität Dortmund (FB Informatik)
01/1993 - 12/1993	wissenschaftlicher Angestellter im BMBF-Projekt "WiNA - Wissensverarbeitung in neuronaler Architektur" an der Philipps-Universität Marburg (FB Mathematik - FG Informatik)
01/1994 - 12/1996	wissenschaftlicher Angestellter des Landes Hessen im FB Mathematik, FG Informatik an der Philipps-Universität Marburg (AG Künstliche Intelligenz und Neuroinformatik)
01/1997 - 12/1997	wissenschaftlicher Angestellter des Landes Hessen im FB Mathematik und Informatik an der Philipps-Universität Marburg (AG Datenbanksysteme)
seit 09/1998	Angestellter der DekaBank (seit 1.1.1999 DGZ-DekaBank) in Frankfurt a.M. im Bereich Organisation und Informatik