Austrian Journal of Statistics
*May 2014, Volume 43, 151–162.*
*http://www.ajs.or.at/*

# Ein Interview mit Wilfried Grossmann

**Wilfried Grossmann**
University of Vienna

**Werner Müller**
Johannes Kepler Univ.

**Matthias Templ**
TU Wien & Statistics Austria

**Abstract**

Das Interview mit Wilfried Grossmann wurde von Werner G. Müller und Matthias Templ am 17.2.2014 durchgeführt. Es beleuchtet das historische Klima des Statistikinstitutes an der Universität Wien, die Ausrichtung der Statistik als „breite“ Datenwissenschaft and der Universität Wien, die Kooperation mit der TU Wien und anderen Institutionen, sowie das Verhältnis zur Statistik Austria, und zwischen der Amtlichen Statistik und der Universitätsstatistik. Zusätzlich wird die Rolle der ÖSG und von EUROSTAT beleuchted. Das Interview widmet sich ausserdem dem Studium der Statistik im Wandel der Zeit - von Lochkarten bis zur Softwareumgebung R und Big Data.

Wilfried Grossmann war Professor für Statistik am Institut für Statistik und danach an der Fakultät für Informatik der Universität Wien Forschungsgruppenleiter der Arbeitsgruppe Data Analysis and Computing. Er hat über 100 Forschungsarbeiten publiziert im Bereich Computational Statistics, Statistisches Datenmanagement, Angewandte Statistik, Theoretische Statistik und Operations Research. Seine aktuellen Forschungsinteressen gelten dem Statistischen Datenmanagement und Informationssystemen, Statistical Computing im Bereich der Amtlichen Statistik, Statistical Knowledge Management, Lehre in der Statistik und Informatik, und Anwendungen von Methoden des Data Mining.

*Keywords*: interview, computational statistics, official statistics.

**Werner Müller:** *Vielen Dank, dass Du dich bereit erklärt hast für dieses Gespräch. Es ist die Idee gekommen in der Österreichischen Zeitschrift für Statistik durch diese Interviewreihe, die hoffentlich eine Reihe wird, auch ein bisschen den Hintergrund unserer Wissenschaft beleuchten zu können. Also nicht immer nur wissenschaftliche Beiträge sondern auch wie es dazukommt, wie man in die Lage kommt, eine wissenschaftliche Karriere in der Statistik überhaupt zu verfolgen in deinem Fall erfolgreich bis zur jetzt erfolgten Pensionierung und hoffentlich auch darüber hinaus. Wie bist du zur Statistik gekommen, bei dir war es ja über die klassische Schiene über die Mathematik?*

**Wilfried Grossmann:** Ich bin eigentlich zufällig zur Statistik gekommen. Als ich mit dem Mathematikstudium begann habe ich von Statistik gar nichts gewusst. Statistik war

damals in der Öffentlichkeit kaum präsent. Ein ehemaliger Schulkollege, der schon ein Jahr vor mir zu studieren begonnen hatte, hat mich mit Erich Neuwirth bekannt gemacht, der ebenfalls Mathematik studierte und als zweites Fach nicht wie traditionell üblich Physik, sondern Statistik im Rahmen eines Studium irregulare. Er erzählte mir von dieser Möglichkeit und ich fand das eine sehr interessante Kombination und entschloss mich ebenfalls diesen Weg zu gehen. Vertreter der Statistik war damals Gerhart Bruckmann, der auf der Wirtschafts- und Sozialwissenschaftlichen Fakultät in Wien neu berufen war. Ich habe diese Entscheidung nicht bereut, weil ich im Laufe des Studiums sehr bald erkannt habe, dass meine Art zu denken wohl eher eine statistische ist als eine mathematische. Mein Studium war natürlich stark mathematisch orientiert und ich habe bei Leopold Schmetterer eine Dissertation über asymptotische Statistik geschrieben. Schmetterer wurde damals als Nachfolger von Swlatscho Sagoroff berufen und wechselte von der Mathematik zur Statistik, also von der Philosophischen auf die Wirtschaft- und Sozialwissenschaftliche Fakultät. Er hat Mitarbeiter für die neue Stelle gesucht und so wurde ich im Jahr 1973 als wissenschaftliche Hilfskraft am damaligen Institut für Statistik angestellt. Ich bin also durch Zufall zur Statistik gekommen, aber ich glaube, dass die damalige Entscheidung für mich die Richtige war.

**Werner Müller:** *Erzähl uns etwas über das historische Klima und die Anfänge dieses Instituts.*

**Wilfried Grossmann:** Was mir gefallen hat und auch meiner Mentalität entspricht war die unerhörte Breite die am Institut für Statistik vertreten war. Ich bin jemand, der immer Interesse hatte Zusammenhänge zwischen verschiedenen Bereichen zu sehen und zu verstehen und nicht so sehr sich in einem engen Bereich zu spezialisieren. Das Klima am damaligen Statistikinstitut war primär von den Persönlichkeiten Gerhart Bruckmann und Leopold Schmetterer geprägt. Bruckmann hatte zwar Mathematik studiert, hat sich aber immer mehr als sozialwissenschaftlicher Statistiker verstanden, während Schmetterer immer Mathematiker und mathematischer Statistiker war. Dann gab es in dieser Zeit an diesem Institut auch den Beginn der Informatik an Österreichischen Universitäten, was mir ganz wichtig erscheint. Es ist interessant, dass die Informatik eigentlich an den Universitäten in Wien aus der Statistik heraus entstanden ist. Swlatscho Sagoroff, der Vorgänger von Schmetterer, hat den ersten Rechner auf einer Österreichischen Universität initiiert. Die Mitarbeiter des Rechenzentrums an der Unversität Wien hatten daher damals eine enge Verbindung mit der Statistik. Eine Person, die in dieser Entwicklung unterschätzt und selten genannt wird, ist Gerhard Derflinger, der dann Statistikprofessor an der Wirtschaftsuniversität wurde. Derflinger hat damals Programme für die Faktorenanalyse erstellt und war weltweit führend in der Entwicklung algorithmischer Lösungen der Faktorenanalyse. Es gab also am Institut die sozialwissenschaftlichen und wirtschaftliche Anwendungen auf der einen Seite, die mathematische Grundlage auf der anderen Seite und natürlich auch die Idee, dass man in Anwendungen die Statistik mit Hilfe der Informatik umsetzen muss. Die Informatiker selbst beschäftigten sich sowohl mit Fragen der Datenorganisation und -speicherung, eine traditionelle Anwendung in der amtlichen Statistik, als auch mit computationalen-algorithmischen Lösungen von statistischen Problemen, das war ein sehr breites Klima.

Bei aller Spezialisierung waren sowohl Schmetterer als auch Bruckmann generell vielseitig interessiert. Bruckmann war der Meinung, dass Statistik das Studium für Generalisten sei, das hat er immer propagiert und in gewissen Sinne hat dieses Institut den Anspruch erhoben, alle Art von Anwendungen formaler Methoden in den Bereichen der Fakultät betreuen zu können. Es gab hier nicht nur die Statistik, es gab Operations Research und Ökonometrie und die Informatik selbst, also die praktische Umsetzung. An diesem Institut herrschte die Vorstellung, „Wir machen das alles“. Im Gegensatz zur TU, da gab es schon sehr früh ein Institut für Operations Research, ein Institut für Ökonometrie, wo sehr gezielt in diesen Bereichen geforscht wurde und natürlich ein

Institut für Statistik und Wahrscheinlichkeitstheorie. Das war für die TU ganz selbstverständlich. Das Institut an der Universität Wien hat für sich immer in Anspruch genommen, das alles mit sehr knappen personellen Ressourcen abzudecken. Es gab 3 Professoren, vielleicht 10 Assistenten und Assistentinnen, und die Unterstützung von externen Lektoren, besonders vom Rechenzentrum, das wie gesagt zu Beginn de facto in Personalunion mit dem Institut für Statistik betrieben wurde. Daher war es so, dass man als Mitarbeiter an diesem Institut nicht nur Statistikvorlesung betreuen musste, sondern auch Mathematik-Lehrveranstaltungen, die Lehre in Operations Research und Ökonometrie, es wurde alles abgedeckt. Das war natürlich sehr interessant und eine Herausforderung.

**Werner Müller:** *Du hast das jetzt angesprochen mit der Informatik, es war ja in Linz auch so ähnlich. Da hat der Kollege Adam das Informatikstudium gegründet, es war wahrscheinlich die Tendenz der Zeit. So gesehen wollen wir darüber sprechen über den Zweig der Statistik der sich wahrscheinlich am stärksten an dieser Entwicklung der computationale Technik entsponnen hat und über die computationale Statistik, wo du eine führende Rolle gespielt hast in den Anfängen. Dann ist da auch die CompStat-Reihe, vielleicht möchtest du auch darüber reden.*

**Wilfried Grossmann:** Nicht nur zur Statistik bin ich durch Zufall gekommen sondern auch viele andere Entscheidungen in meinem Leben sind durch äußere Einflüsse zufällig zustande gekommen. Wenn mich etwas interessierte und der Meinung war, das könnte man sich ansehen und etwas machen, versuchte ich es umzusetzen. So war das auch bei der Entstehung von CompStat. Als ich im Jahr 1974 das Studium abgeschlossen hatte und fix angestellt wurde, gab es im Herbst 1974 den ersten CompStat Kongress. Die Hauptinitiatoren dieses Kongresses waren Peter Paul Sint und Johannes Gordesch, die haben das Ganze initiiert. Wohl in der Tradition, dass an diesem Institut immer computationale Statistik betrieben wurde, insbesondere Anwendungen in der Psychologie, der Chemie und der Physik, hatten sie die Idee einen Kongress zu organisieren. Gordesch ist kurz vor dem Kongress nach Berlin berufen worden, und Sint musste alleine die Organisation übernehmen. Peter Paul Sint ist ein hochintelligenter und unerhört vielseitig interessierter Mann, aber allein war die Organisation für ihn nicht zu schaffen. Da die zentrale Postadresse des Kongresses das Institut war und Sint selbst nicht mehr am Institut beschäftigt war, haben Georg Pflug und ich begonnen Sint bei der Organisation von CompStat zu unterstützen. Der Kongress war dann auch ein Erfolg und ist heute noch eine wesentliche wissenschaftliche Veranstaltung in der computationalen Statistik. Dadurch ist meine Nähe zu der computationalen Statistik gekommen und ist auch ein wesentliches wissenschaftliches Interesse für mich geblieben. Anfangs habe ich mich für algorithmische Fragen interessiert. Das war die algorithmische Lösung für nichtlineare Regression, also eher numerische Probleme. Da am Institut damals die Informatik immer mehr an Bedeutung gewonnen hat, haben wir uns im Laufe der Zeit für Fragen der statistischen Softwareentwicklung, der Simulation, der statistischen Expertensystemen und Fragen der statistischen Datenorganisation interessiert. Dadurch hat es immer eine gewisse Verbundenheit mit den CompStat Entwicklungen gegeben und auch ein Interesse, dass diese Idee weiter Bestand hat. Wir haben dann gemeinsam mit Kollegen Rudi Dutter von der TU im Jahr 1994, den Kongress anlässlich 20 Jahre CompStat an der TU organisiert. Bei diesem Kongress ist meines Wissens nach erstmals SPlus® im wissenschaftlichen Programm und als Aussteller verstärkt aufgetreten. SPlus® wurde ja dann bald von R abgelöst und die heute dominante Rolle von R in der Statistik ist sicher zu einem guten Teil der Verdienst von Kurt Hornik und Friedrich Leisch. Ihre Leistungen für die computationale Statistik ist weit bedeutender als meine eigene.

Weil du das angesprochen hast, möchte ich vielleicht noch einen Punkt über die Verbindung zwischen Statistik und Informatik ansprechen, der aus heutiger Sicht wahrscheinlich eine Fehleinschätzung war. Das ist die Einführung und Planung der Wirtschaftsin-

formatik die an der Universität besonders von Bruckmann sehr stark forciert wurde. Das Engagement für die Wirtschaftsinformatik war sicher richtig auf Grund der Bedeutung und Wichtigkeit der Informatik. Man hat bei der Planung der Wirtschaftsinformatik aber übersehen, dass man im Studium die computationale Statistik als einen essentiellen Bestandteil des Studiums etabliert. Man hat anfangs eine Betriebsinformatik mit Operations Research und eine Wirtschaftsinformatik mit Ökonometrie eingerichtet, das war klar. Aber dass natürlich die Statistik selbst für beide Zweige ein essentieller Bestandteil ist, hat man zu wenig beachtet. Aus heutiger Sicht ist das sicher nicht richtig gewesen, insbesondere wenn man an die heutige Bedeutung von Data Mining im Zusammenhang mit Business Intelligence denkt. Durch die Wirtschaftsinformatik ist das Statistikstudium für Studierende ein bisschen uninteressanter geworden und in eine Nische gewandert. Viele sahen die Statistik primär als eine Verwaltungswissenschaft zur Unterstützung der BWL und der Ökonomie. Erst als du 1983 studiert hast ist es etwas besser geworden, weil sich immer mehr Anwendungen der Statistik ergeben haben. Aus heutiger Sicht würde ich die damalige Entscheidung als Planungsfehler bezeichnen. Aber man kann nicht alles im Vorhinein wissen. Bruckmann hat, wie damals wohl weltweit die meisten Wissenschaftler, mehr auf Operations Research als wesentliches Planungsinstrument für Management Science gesetzt und nicht auf eine datenzentrierte Management Science, die heute als Business Intelligence im Vordergrund steht.

**Matthias Templ:** *Für die jungen Leser des AJS. Wie wurde damals gerechnet? Wenn man z.B. nichtlineare Regression denkt; es war irrer Aufwand dies zu implementieren, sich die LAPACK-Routinen zu besorgen.*

**Wilfried Grossmann:** Genauso war es. Es gab ein Rechenzentrum an der Universität Wien, das war im Keller des neuen Institutsgebäudes, da gab es einen großen Raum, mit Lochkartendruckern.

Da hat man die Jobs gestanzt. Dann hat man die fertigen Jobs fertig in den Kartenleser gesteckt, die sind gelesen und verarbeitet worden, und dann musste man warten, bis auf diesem Endlospapier ein Output gekommen ist. Wenn man einen Fehler gemacht hat, hat man wieder von vorne begonnen. Ein schnelles effektives Arbeiten in heutigem Sinn war undenkbar, dafür war es kommunikativ. Wenn man gewartet hat bis das Programm fertig ist, hat man draußen mit anderen Leuten getratscht, es waren hauptsächlich Chemiker, Physiker und Psychologen dort. Es war alles eine Batchverarbeitung. Meist waren das Fortran-Programme, die Libraries wie die NAG Library für das numerische Rechnen verwendet haben, später kam dann erst PASCAL. Als wir uns dann mit Verkehrssimulation beschäftigten sind auch noch andere Sprachen wie SIMULA dazu gekommen.

Langsam setzten sich erst Terminals durch, wo man Plätze reservieren musste. Dass jeder einen eigenen Rechner mit einer geeigneten Arbeitsumgebung hat, wie es heute der Fall ist, war unvorstellbar. Auch der Speicherplatz und die Rechenkapazität waren begrenzt. Als ich in den späten 80er Jahren und frühen 90er Jahren an Projekten zur Ökosystemsimulation arbeitete haben wir Ozonkonzentrationen analysiert. Wenn ich da eine Clusteranalyse für die Tagesgänge machen wollte musste ich extra eine Erweiterung des Speicherplatzes anfordern und man musste die Daten selbständig partitionieren. Problemlösungen für heute selbstverständliche Datenvolumina waren oft illusorisch.

**Werner Müller:** *Weil du es vorher schon angesprochen hast: die amtliche Statistik. Es gibt ja ein gesundes Spannungsverhältnis zwischen der amtlichen und der akademischen Statistik. Vielleicht kannst du uns darüber etwas erzählen, wie sich das im Laufe der Zeit*

*entwickelt hat.*

**Wilfried Grossmann:** Das ist ein interessanter Punkt und bin über die derzeitige Entwicklung sehr froh, wenn ich sie mit der Vergangenheit vergleiche, als noch Leopold Schmetterer und Lothar Bosse die Vorsitzenden der Statistischen Gesellschaft waren. Ich kann mich an eine Gespräch mit Bosse bei einer Veranstaltung erinnern, als er sagte, es gibt zwei Bereiche in der Statistik, das eine ist die mathematische Statistik das andere ist die amtliche Statistik, das sind zwei Welten die wenig miteinander zu tun haben. Damals war das auch so, dass diese zwei Bereiche eher getrennt voneinander in der Statistischen Gesellschaft lebten. Jede Gruppe respektierte die andere, aber es war mehr ein „teile und herrsche Prinzip". Bei Veranstaltungen trat man gemeinsam auf und Veranstaltungen einer der beiden Gruppen wurde immer finanziell gefördert. Für die universitäre Statistik war dies auch eine große Hilfe, da die Einnahmen der Gesellschaft zu einem überwiegenden Teil aus dem Bereich der amtlichen Statistik kamen. Wissenschaftlich war die amtliche Statistik damals sicher näher an der Informatik, insbesondere Fragen Datenorganisation und Datenspeicherung spielten eine zentrale Rolle, da sie klarerweise sehr eng mit den Aufgaben amtlichen Statistik, die ja statistische Information verwalten und bereitstellen muss, in Verbindung stehen. Hier muss man Präsident Bosse im Nachhinein große Weitsicht zugestehen und auch und großes Lob aussprechen. Er erkannte sehr früh, dass in der amtlichen Statistik statistische Datenbanken eine zentrale Rolle spielen und er hat dementsprechend die Entwicklung im statistischen Zentralamt sehr gefördert und eine Gruppe unter Lutz arbeiten lassen. Diese Gruppe hat ein sehr fortschrittliches Modell für statistische Datenbanken entwickelt, es war damals weltweit eines der besten Systeme. Die Ideen wurden dann von anderen Ländern aufgegriffen, in Österreich ist die Entwicklung leider etwas eingeschlafen.

An methodischer Statistik gab es in der amtlichen Statistik nur geringes Interesse und beschränkte sich auf einfache summarische Statistiken wie Summe, Mittelwerte, Indizes und eventuell noch Varianz. Dazu hat man einfache Grafiken gemacht, mehr gab es nicht. Daher gab es immer ein gewisses Spannungsfeld, das durch das persönliche Geschick und die Toleranz der beiden Vorsitzenden, insbesondere Schmetterer und Bosse aber auch danach Bruckmann, Josef Schmidl und Reinhold Viertl, ausgeglichen werden konnte. Dass man in der amtlichen Statistik auch komplexere Methoden verwendet war damals unüblich. Die wesentliche Methode der amtlichen Statistik war die Stichprobenziehung. Die Stichprobentheorie wurde ja vom Neyman als Grundlage für alle statistischen Verfahren entwickelt und das Prinzip der Randomisierung spielt ja auch für statistische Modellen eine zentrale Rolle. Dann hat sich aber die methodische Statistik rasch weiter entwickelt und in den österreichischen Statistikcurricula ist amtliche Statistik kaum vorgekommen. Im Statistikstudium standen statistische Modelle und deren Fundierung durch die Wahrscheinlichkeitstheorie im Vordergrund.

Stichprobentheorie war nur ein Thema am Rande, mit Ausnahme der Biometrie, aber die Versuchspläne für statistische Experimente setzen doch einen anderen Schwerpunkt als die amtliche Statistik. Modelle, die für die amtliche Statistik interessant sind, sind erst später entwickelt worden, aber da hat es in Österreich kaum Beiträge gegeben. In Österreich waren es zwei Welten. In anderen Ländern haben Statistiker mit Modellen für Modell Assisted Survey Information Collection, oder Model Based Survey Information neue Impulse gesetzt. Besonders in England, Schweden und den USA und Canada wurde diese Entwicklung vorangetrieben. Das begann Ende der 70-iger Jahre. Es gab dann auch die Diskussion über die den Design Based und den Model Based Zugang zur Statistik. Ein Meilenstein für die amtliche Statistik und auch für die computationale Statistik ist meiner Meinung nach die Entwicklung des EM-Algorithmus 1977 durch Dempster Laird und Rubin und dessen Anwendung zur Behandlung von fehlenden Werten.

Heute gibt es erfreulicherweise eine Vielzahl von Methoden die ganz wichtig für die amtliche Statistik sind. Auszählen und summarische Statistiken allein sind nicht mehr

ausreichend. Daten müssen sorgfältig vorverarbeitet werde, z.B. bei fehlenden Werten, damit die Ergebnisstatistiken eine entsprechende Qualität haben. Aber auch für die Modellierung von Zusammenhängen von Daten in der amtlichen Statistik spielen heute Modelle eine größere Rolle. Die Glättung von Zeitreihen ist ein klassisches Beispiel, neuere Anwendungen sind Small Area Estimation um globale Zahlen auf lokale Ebene umzulegen. Es gibt also heute eine Reihe von Bereichen der methodischen Statistik, die für die amtliche Statistik interessant sind. Heute ist die amtliche Statistik mehr als Datenverarbeitung und nicht mehr von der methodische Statistik zu trennen. Das wird auch in den Themen der NTTS-Konferenzen (New Techniques and Technologies for Official Statistics) deutlich, die von EUROSTAT initiiert wurden. Anfangs standen Fragen der Datenorganisation und Metadaten im Vordergrund, heute findet man mehr Beiträge zur statistischen Methodik. Ich empfinde das als eine positive Entwicklung, weil sie zeigt, dass man zu Recht von einer Statistik sprechen kann.

**Werner Müller:** *Dass das Verhältnis von Statistik Austria und der akademischen Statistik nicht vollständig auseinandergedriftet ist in den 70-iger und 80-igern war vielleicht ein Verdienst der ÖSG und der Protagonisten damals.*

**Wilfried Grossmann:** Ja, das ist zweifelsohne ein Verdienst der ÖSG. Weil die ÖSG mit dieser Konstruktion der Doppelvorsitzenden (eine Person aus dem amtlichen Bereich, eine Person aus dem universitärem Bereich) immer versucht hat die Balance aufrecht zu halten. Man war sich bewusst, dass es schwierig ist eine gemeinsame Sprache zu entwickeln aber beide Seiten haben einander respektiert, das ist ein ganz wichtiger Punkt. Innerhalb der ÖSG war immer ein Respekt der beiden Bereiche in einem hohen Maße vorhanden. Besonders erwähnen möchte ich in diesem Zusammenhang Alfred Franz, der sich in seiner Zeit als Sekretär der Gesellschaft sehr um eine stärkere wissenschaftliche Ausrichtung bemüht hat und an der methodischen Statistik immer sehr interessiert war. Die ÖSG hat viel zu einem gemeinsamen Verständnis der Statistik beigetragen. Als es dann Ende der 80-iger Jahre zu Schwierigkeiten im gegenseitigen Verständnis kam wurde die Gesellschaft ja neu strukturiert. Das kann man in dem Artikel über die Geschichte der ÖSG, auf der Homepage nachlesen. Mitte der 90-iger Jahre wurde die Gesellschaft unter der Leitung von Peter Hackl reorganisiert und das neue Modell ist seither sehr erfolgreich. Präsident Joachim Lamel hat das Motto, Triple A ausgegeben: Amtliche Statistik, Angewandte Statistik und Akademische Statistik. Die Statistische Gesellschaft deckt alle diese Bereiche ab und ein auseinanderdriften muss verhindert werden. Das scheint mir sehr wichtig und scheint auch in Österreich gut gelungen.

**Matthias Templ:** *Vielleicht ist doch das ein bisschen zu kritisieren bezgl. den Universitäten in Österreich. Du hast Stichwörter genannt wie Rubin, missing values, design-basierte Verfahren, model-assisted Verfahren. Mit Ausnahme von Linz, wo Quatember und Bacher diesen Bereich betreuen, wird das in Österreich nicht mehr gelehrt. Die ganze Problematik der Methoden in der Offiziellen Statistik wird weitgehend nicht behandelt. Wenn man andere Ländern vergleicht, wie z.B. Deutschland, gibt es sehr wohl Lehrstühle welche die Methoden der Offiziellen Statistik und Stichprobentheorie abdecken. Siehst du das befremdlich?*

**Wilfried Grossmann:** Das Problem sehe ich, aber es ist vielleicht eine Art von Pendelbewegung, wenn man die Geschichte des Statistikstudiums an der Universität Wien ansieht. Am Beginn war das ein sozial- und wirtschaftswissenschaftliches Studium, der Anteil der methodischen Statistik im Studium war eher bescheiden. Es gab anfangs Lehrveranstaltungen aus der amtlichen Statistik. Mitte der 80-iger Jahre gab es die erste Studienreform, die stärker den methodischen Bereich der Statistik in das Statistikstudium einbringen sollte, ich war auch der Meinung, dass das notwendig war. Gleichzeitig wollte man die Bereiche Volkswirtschaft und die Betriebswirtschaft reduzieren. Werner, hast du noch Buchhaltung lernen müssen?

**Werner Müller:** *Natürlich, ich bin noch aus dieser Vorgeneration.*

**Wilfried Grossmann:** Eben, es war ein wirtschaftswissenschaftliches Studium. Buchhaltung und Kostenrechnung waren essentiell, weil man der Meinung war, das müssen die Absolventen können. Mit der Reduktion der Wirtschaftswissenschaften ist auch die amtliche Statistik, die ja einen starken Bezug zur Volkswirtschaft hat, im Studienplan reduziert worden. Die Lehrveranstaltung für Amtliche Statistik hat dann Alfred Franz betreut. Ich habe auch einige Male gemeinsam mit Hofrat Franz eine Lehrveranstaltung zum Thema Amtliche Statistik gemacht. Dann habe ich gemeinsam mit Karl Fröschl und Marcus Hudec diesen Bereich betreut. Wir haben uns in Kenntnis der internationalen Entwicklung bemüht die Verwendung von Methoden in der amtlichen Statistik in das Vorlesungskonzept einzubeziehen. In der letzten Studienreform ist die amtliche Statistik ganz rausgefallen. Leider, denn ich sehe amtliche Statistik als eine wichtige und spezifische Anwendung der Statistik mit eigenständigen Fragestellungen.

Es gab dann noch im Informatikstudium den Zweig „Data Engineering and Statistics“, da gab es auch eine Lehrveranstaltung zur Amtlichen Statistik. Leider gibt es jetzt das Studium nicht mehr, da es sich nicht durchgesetzt hat, vermutlich weil die Organisation im Rahmen eines Informatikstudiums schwierig war.

Aber das ist sicherlich ein Fehler, dass man bei der Planung und Adaption der Statistikstudien, wie ich vorher gesagt habe, jene internationalen Entwicklungen nicht berücksichtigt hat, die helfen die Kluft zwischen amtlicher Statistik und methodischer Statistik zu schließen. Ich persönlich sehe die amtliche Statistik als einen speziellen Bereich der angewandten Statistik. Amtliche Statistik ist eine spezielle Anwendung der Statistik, traditionell in Verbindung mit den Wirtschaftswissenschaften, insbesondere in der Makroökonomie und den Sozialwissenschaften. Der Schwerpunkt war historisch die Bereitstellung der Daten, es ist also eine spezielle Anwendung. Und jede spezielle Anwendung der Statistik erfordert spezielle Methoden, wir haben über die schon vorhin gesprochen. Es ist im Grunde genommen ähnlich zu sehen, wie andere Anwendungen, z.B. im Marketing, in der Betriebswirtschaft, in der Technik oder in der Medizin. Natürlich spielen in der Medizin andere Fragen eine Rolle. Andere Modelle, die scheinbar nichts mit Amtlicher Statistik zu tun haben wurden aber in letzter Zeit interessant. Matthias, Du hast ja in dieser Richtung viel gemacht, z.B. die robuste Statistik ist für viele Fragen der amtlichen Statistik interessant. Nicht im Sinne des Outlier-Modells, dass man fehlerhafte Daten hat, sondern dass man kleine exzeptionelle Gruppen hat, das ist ein essentielles Problem in der Wirtschaftsstatistik, und die müssen geeignet berücksichtigt und dargestellt werden. Es ist auch interessant dass es im Bereich der Registerzählung methodische Anwendungen gibt. Wir haben kürzlich mit Kollegin Lenk eine Anwendung von Klassifikationsverfahren für die Frage der Zuordnung von Personen zu bestimmten Gruppen gemacht, wenn man diese Information fehlt. Wir haben logistische Regression angewendet und auch Boosting, um von Personen festzustellen ob sie in Österreich wohnhaft sind oder nicht. Eine weitere Anwendung sind Belief-functions um zum Beispiel den plausibelsten Familienstand zu bestimmen, wenn in verschiedenen Registern unterschiedliche Einträge sind.

Es ist also heute nicht mehr so, dass der Bereich der Methoden in der Amtlichen Statistik auf wenige einfache Verfahren beschränkt ist. Auf der anderen Seite wäre es vielfach vorteilhaft, wenn die methodischen Statistiker in ihren Anwendungen die Genauigkeit der amtlichen Statistik hinsichtlich der Dokumentation der Datenerhebung übernehmen würden. Das gilt für viele Anwendungen, die unter dem Titel Data Mining gemacht werden, insbesondere wenn Daten vom Internet verwendet werden. Es wäre nicht schlecht, sich zu fragen, woher diese Zahlen kommen, wie valide sie sind, und wie die Grundgesamtheit aussieht die sie repräsentieren sollen. Vielfach glaubt man solche Daten seien eine Vollerhebung, dabei weiß man gar nicht wie die Grundgesamtheit aussieht. Also diese ganzen klassischen Fragen der Datenqualität die in der amtlichen Statistik zentral

sind, die sollte man sich auch bei der Anwendung der Statistik in anderen Bereichen immer wieder stellen.

**Matthias Templ:** *Big Data zum Beispiel...*

**Wilfried Grossmann:** Gerade bei Big Data sind solche Fragen eine Herausforderung. Ich habe von Ralf Münnich bei den letzten Statistiktagen gehört, dass es jetzt bei EUROSTAT einen Arbeitskreis gibt, der sich genau mit diesen Fragen beschäftigt. Eine Lösung des Problems ist sicher nicht einfach und erfordert gute Ideen.

**Werner Müller:** *Es soll ein Master für Offical Statistics, European Master kreiert werden, darüber wird schon längere Zeit gesprochen. In Linz, in unseren Studienplänen ist die Amtliche Statistik sehr wohl noch vorhanden.*

**Wilfried Grossmann:** Das ist sicher ein möglicher Schwerpunkt und das würde ich positiv sehen. Wenn ich heute die Statistikstudien in Österreich ansehe, haben sie sich generell sehr positiv entwickelt. Der Aufbau des Studiums in Linz, was dort den Studierenden geboten wird, das halte ich für sehr gut, das gefällt mir sehr gut. Es trifft vielleicht am besten meine heutige Sicht zur Statistik die eben von der Vorstellung geprägt ist, dass es eine Einheit gibt zwischen den unterschiedlichen Bereichen und das diese gemeinsam behandelt werden sollten. In Wien sind etwas andere Schwerpunkte und die Entwicklung der Studierendenzahlen ist sehr gut. Auch an der WU hat sich die Statistik sehr gut entwickelt und der Bereich der Statistik ist jetzt dort stark vertreten und sehr aktiv. In Salzburg ist jetzt mit Kollegen Arne Bathke auch ein neuer Schwung in die Statistikausbildung gekommen. Dort ist die Statistik soweit ich weiß enger mit der Mathematikausbildung verbunden und es ist wichtig auch in diesem Bereich präsent zu sein.

Weil ihr vorhin gefragt habt, wie ich zur Statistik gekommen bin. Heute ist es ja ganz anders, weil Statistik in aller Munde ist und auch der Stellenmarkt ist viel besser. Früher war der Stellenmarkt doch sehr beschränkt auf den Verwaltungsbereich oder den Akademischen Bereich. Aber heute finden StatistikerInnen unter dem Titel Data Scientist in der Wirtschaft, bei Versicherungen oder bei Banken sehr gute Berufschancen. Es ist positiv, dass es so viele Möglichkeiten zur Spezialisierung gibt. Es gibt ja auch den riesigen Bereich der Bioinformatik, der in Wien ja durch Andreas Futschik vertreten war. Da sind ja völlig neue Anwendungsfelder gekommen, die nicht Statistik genannt werden sondern Bioinformatik, Machine Learning oder Data Mining. Die Statistik ist oft nicht so clever, dass sie ihren Beitrag richtig vermarkten kann, vielleicht auch, weil sie traditionell ein bisschen akribischer mit der Qualität der Information umgeht, dass muss man auch positiv sehen. Ich sehe es nicht negativ, dass man bei der Vermarktung etwas ins Hintertreffen kommt, aber generell sind die Möglichkeiten sehr gut.

Und der Master in Official Statistics, das ist eine neue Spezialisierung in einem interessanten Bereich, weil die internationalen Organisationen in ihren Analysen sehr viel statistische Modellierung verwenden. Man darf nicht übersehen, dass die PISA Studie ein hochkomplexes statistisches Modell ist mit einem psychometrischen Modell im Hintergrund. Die Wirklichkeit ist von der Idee, da fragen und zählen wir, wieviel richtig und wieviel falsch, meilenweit entfernt. Es steckt soviel statistische Methodik dahinter, dass zum Verständnis des Details eine fundierte Statistikausbildung notwendig ist.

**Matthias Templ:** *Darf ich etwas außerhalb des Protokolls fragen oder ist es zu brisant? Die Analyse der PISA Studie war in deinen und Erich Neuwirth's Händen und ich glaube auch Fritz Leisch hat etwas beigetragen, danach wurde der Auftrag aus wenig nachvollziehbaren Gründen anderweitig vergeben. Ich habe dann verwundert weniger professionelle Vorträge gehört, und z.B bekreidet dass selbst die Problematik der fehlenden Werte nicht berücksichtigt wurde. Ist dass ein Politikum geworden?*

**Wilfried Grossmann:** PISA ist leider ein Politikum geworden. Wir haben ja etwas auf Initiative von Erich Neuwirth gemacht. Das ist im Zusammenhang mit PISA 2003 zustande gekommen. Da hat es die große Aufregung gegeben hat, weil die Österreicher plötzlich so schlecht abgeschnitten haben. Da war besonders Erich Neuwirth aktiv und hat begonnen sich die Daten herunterzuholen und anzusehen. Dann haben wir begonnen zu diskutieren und haben uns die ganzen Unterlagen angesehen. Wir haben versucht das Modell zu verstehen und nachzuvollziehen. Da haben wir festgestellt, dass der Datenerhebungsaspekt und der psychometrische Aspekt nur ein Teil sind. Von der Psychologie hat ja auch Ivo Ponocny mitgearbeitet. Das Modell für die Skalierung der Items ist ein psychometrisches Modell, dass auf dem Rasch Modell beruht. Wenn man das statistisch sieht ist es in Wirklichkeit ein multivariates logistisches Regressionsmodell, also ein komplexes Generalized Linear Model. Und das Ganze wird dann noch eingebettet in einen Bayesianischen Kontext um die Effekte der Schulen und den individuellen Effekt zu berücksichtigen. Und dann wird aus diesem Modell mit Methoden der Analyse von fehlenden Werten eine Vorhersage der Scores mit Hilfe von multiplen Imputation gemacht. Da stecken also in Wirklichkeit sehr viele Modelle drin und die Schätzung der Varianzen ist dann noch ein eigenes Problem, aber da kennst du dich besser aus mit der Methode von Fay zur Varianzschätzung.

Wir haben also die Struktur des Modells analysiert und die einzelnen Bestandteile im Detail angesehen. Dabei haben wir festgestellt, dass bei der Berechnung der Scores, defacto sind die Scores die man erhält Posterior-Mittelwerte, eine gewisse Ungenauigkeit ist. Wir haben dann eine Verbesserung vorgeschlagen, die auch von der PISA akzeptiert wurde, da waren wir interessiert dran. Es ist nicht leicht, das Modell statistisch zu analysieren. Daher gibt es immer wieder Aufträge, Andreas Quatember hat jetzt wieder etwas gemacht, es gibt so viele Aspekte. Du kannst die Stichprobenkonzeption hinterfragen, man kann das Rechenmodell hinterfragen, man kann psychometrische Kalibrierung hinterfragen, es gibt so viele Komponenten die alle hinterfragbar sind. Ob das Ministerium das will, ist wiederum eine andere Frage, aber vom statistischen Standpunkt ist das das Interessante bei der ganzen Sache. Ich glaube PISA ist ein sehr gutes Beispiel, wo heute Survey Statistik hingeht. Es reicht nicht mehr aus, dass du einen Fragebogen ausgibst und dann zählst wieviel Angestellte ein Unternehmen hat.

**Werner Müller:** *Diese Schilderung scheint mir auch eine gute Illustration der Entwicklung der Statistik im Allgemeinen zu sein: dass immer mehr Layer kreiert werden. Auch in der Methodik und dass das dann an und für sich relativ undurchschaubar wird, für jemanden der das vielleicht analysieren muss. Es gibt ja schon Bereiche, wo man wieder dazu übergeht Metamodelle zu bauen, weil man die eigentlichen Modelle nicht mehr versteht. Hast du da Gedanken zu dieser Entwicklung?*

**Wilfried Grossmann:** Was du ansprichst, ist eine sehr schwierige Frage. Ich glaube man muss jedes Modell immer in seiner praktischen Anwendbarkeit hinterfragen. Man muss sich fragen, was gewinne ich durch das komplexe Modell oder wird durch das komplexe Modell in Wirklichkeit alles noch schwieriger zu interpretieren. Diese komplexen Modelle kommen oft dadurch zustande, dass wir immer mehr Informationen haben und wir wollen in einem Modell alle verfügbaren Informationen reinrechnen. Wir machen jetzt ein Projekt gemeinsam mit der Medizinuni, wo es um evidenzbasierte Medizin geht. Da ist die Hoffnung vieler Leute, dass durch ein komplettes Monitoring mehr Information zur Verfügung steht und dieses mehr an Information könnte dann zu besseren und richtigeren Entscheidungen führen. Wir haben eine kleine Diskussionsrunde gehabt, da hat Georg Heinze von der MedUni sehr klar und schön argumentiert, dass im Fall der klassischen Biometrie durch mehr Kovariaten und mehr Confounder die Lage für eine rationale Beurteilung immer schwieriger wird, weil da natürlich oft keine kontrollierten Experimente mehr möglich sind. Es gibt dann nur mehr ganz wenige Fälle und die Interaktion zwischen den einzelnen Confounder sind schwer abzuschätzen. Das Prinzip

eines Parsimonious Models ist auch heute noch wichtig. So gesehen ist Big Data nicht nur ein Segen sondern auch ein Fluch. Big Data ist ein Fluch, weil man nie weiß welche Information man für ein Modell selektieren soll. Zu all der Information kommt noch wahnsinnig viel zeitliche Information dazu. Es gibt ja nur mehr Zeitreihen aber die Zeitreihen von logfiles muss man sehr wohl hinterfragen und nachdenken darüber was man damit machen kann. Leichter wird es dadurch nicht. Die inhaltliche Beurteilung ist ein zentraler Punkt und da ist meiner Meinung nach die Statistik besser aufgestellt als die Informatik. Die Statistik ist es gewohnt für die Daten auch eine inhaltliche Beschreibung zu geben und inhaltlich darüber nachzudenken, da hat die Statistik einen Vorsprung gegenüber anderen Wissenschaften.

**Matthias Templ:** *Wie ich gelesen habe bist noch involviert bei EUROSTAT. (Anm.: jetzt nicht mehr.) Meiner Meinung wird EUROSTAT in letzter Zeit immer mehr zu einer Verwaltung, immer mehr Administration steht im Vordergrund. Da keine Statistiker mehr dort sitzen passiert es z.B. auch bei Big Data das Consultingfirmen beauftragt werden welche das Thema sehr puschen.*

**Wilfried Grossmann:** Big Data und EUROSTAT ist ein schwieriger Punkt. Wir haben voriges Jahr bei den Statistiktagen diese Sektion Big Data gehabt, du wolltest ja dass ich das organisiere. Es war für mich klar, man sucht Vortragende die Big Data aus Sicht des Data Mining repräsentieren. Aber dann habe ich mir gedacht, man sollte auch jemand einladen der Big Data aus der Sicht der Amtlichen Statistik präsentiert und habe mir angesehen, was es im Bereich von EUROSTAT in Big Data gibt. Die meisten Projekte werden nicht an Statistikinstitute oder an StatistikexpertInnen vergeben, sondern die gehen alle an Informatikfakultäten. Für die ist Big Data primär die Frage, wie kann ich die Daten manipulieren, wie kann ich Big Data managen. Zentral ist dass die Daten in ein System eingepackt werden, in Ontologien, xml-Schemata oder so etwas und das muss dann schnellsten verarbeitet werden. Inhaltlich interessieren die Daten fast gar nicht. Das ist sicherlich ein Problem bei der ganzen Sache und EUROSTSAT beschränkt sich hier zu stark auf die Rolle der Verwaltung. Es wird dann nicht mehr kritisch hinterfragt woher die Zahlen kommen, dieser Hype ist eine große Herausforderung, weil man muss das für jeden einzelnen Fall entscheiden. Das Problem ist etwas anders als bei Daten von physikalischen Messungen durch Satelliten oder Genomdatenbanken in der Bioinformatik. Die Probleme der Speicherung und Datenhaltung sind natürlich ähnlich, aber hier gibt es von der Substanzwissenschaft klarere Vorstellungen was gespeichert werden soll und wie diese Information weiter verarbeiten soll. Im Sozialbereich und in der Wirtschaft ist dies oft viel schwieriger, besonders bei Daten über den Internetverkehr. Auf Vorrat nur zu organisieren und das vom rein informatischen Standpunkt her zu betrachten löst zwar das Verwaltungsproblem aber nicht das inhaltliche Problem. Es ist meiner Meinung nach ein Problem, dass die Statistik in diesem Schema derzeit zu kurz kommt, aber wie schon gesagt es besteht die Hoffnung dass sich hier in nächster Zeit etwas bewegen wird.

**Werner Müller:** *Vielleicht sind wir eh schon an dem Punkt, wo wir uns ein bisschen über die Zukunft unterhalten können. Es ist nichts schwerer, als die die Zukunft vorherzusagen, aber hast du irgendwelche Perspektiven, wie du glaubst, wie sich die Statistik entwickeln wird?*

**Wilfried Grossmann:** Ich glaube die angewandte Statistik und die computationale Statistik werden in der Zukunft weiterhin sehr wichtig sein. Die Rolle der Statistik in der Gesellschaft ist derzeit sehr gut und Statistik gilt als eine sehr wichtige Wissenschaft. Ich glaube nicht, dass die Statistik diese Rolle in nächster Zeit verlieren wird. Statistik ist die Wissenschaft, die Information vom inhaltlichen Standpunkt her am besten verarbeiten kann, aber es wird wohl kaum eine Statistik ohne die Informatik geben. Viele Fragen der theoretischen Statistik sind heute schon im Zusammenhang mit Informatik

zu sehen. Die Handhabung von neuen und komplexeren Datenstrukturen ist dabei ganz wesentlich. Für mich ist das schon schwierig, das sage ich ganz offen, aber für die jungen StatistikerInnen ist das Manipulieren von Big Data und komplexer Anwendungssoftware kein Problem mehr, die beherrschen das ja, weil sie Digital Natives sind. Besonders für eine intelligente Darstellung der Daten wird man auch in Zukunft viel Statistik brauchen. Die Visualisierung wird eine zentrale Rolle spielen, es wird auch die Statistik eine große Rolle spielen. Ein Bereich wo noch viel zu machen ist, ist wie man mit Textdaten umgeht. Soweit ich das verstanden habe ist bei Text Mining die methodische Komponente noch nicht sehr entwickelt. Ich glaube sehr wohl, dass man hier noch eine Menge erreichen kann, nur hat man die Modelle noch nicht ganz im Griff. Nur die Häufigkeiten auszuzählen und diese dann als topic maps darzustellen ist sicher erst ein Anfang, da wird man noch viel Statistik machen können, vielleicht eine andere Art. Das ist noch am Stand wie die amtliche Statistik früher war, es geht nur um Häufigkeiten, aber a la longue wird man auch hier mit statistischen Methoden mehr zustande bringen.

Es gibt viele andere Bereiche die derzeit ein Randgebiet der Statistik sind, zum Beispiel Bildverarbeitung und automatische Übersetzung. Es ist interessant, dass viele Übersetzungsprogramme jetzt eher statistisch orientiert sind. Sie suchen nach statistischen Ähnlichkeitsstrukturen in den sprachlichen Konstrukten. Das wird nur von wenigen Statistikern gemacht und spielt in den Curricula nur am Rande eine Rolle. Aber das sind sicherlich Bereiche wo Statistik sehr wichtige Beiträge leisten kann. Auch dynamische Grafiken werden immer bedeutender und vielleicht sollte sich die Ausbildung für solche Bereiche öffnen. Das wird sicherlich in Zukunft ein wichtiger Markt sein und solche Anwendungen sollten von der Statistik nicht außer Acht gelassen werden.

In der Zeitschrift „Significance“ gab es kürzlich einen interessanten Artikel, wo „Dr. Fisher“, schreibt, irgendwie ist es erstaunlich, das die statistische Methodik, die ja eine Methodik für Datensätze der Größenordnung von 100 Beobachtungen ist und mit einem Taschenrechner betrieben werden kann, auch für die neuen Probleme angewendet werden kann. Die Grundlagen dieser Methodik funktionieren also auch bei großen Datensätzen und werden immer angewendet, es hat sich in diesem Sinne ja nicht viel geändert. Auch in der Bioinformatik verwendet man klassische Methoden. Also die Kunst wird wahrscheinlich wirklich sein, wie man Strukturierung in diese großen Datenmengen mit statistischen Methoden erzeugt. Vielleicht wird man dann zu dem Ergebnis kommen, dass weniger Information nützlicher ist, das ist der Erfolg der Statistik. Dieser Erfolg der Statistik beruht auf statistischen Prinzipien wie der Randomisierung, dem Likelihoodprinzip, oder statistischer Modellierung. Sie erlauben eine Verdichtung der Information die inhaltlich interpretiert werden kann. Und wie diese Prinzipien für diese neue Datenwelt vernünftig umgesetzt werden, das halte ich für eine interessante Frage.

**Werner Müller:** *Was interessiert dich noch, in welcher Rolle werden wir dich noch sehen?*

**Wilfried Grossmann:** Naja, ich habe ja in verschiedenen Bereichen Statistik betrieben, weil ich durch die anfangs genannte Vorstellung, dass Statistik eine Generalisten-Disziplin ist geprägt wurde und ich mir immer wieder gesagt habe, wenn etwas interessantes Neues kommt dann soll man sich damit auseinander setzen. Jetzt ist der Bereich Business Intelligence dazu gekommen, das interessiert mich im Moment. Da sehe ich auch, dass hier vieles von der Informatik gemacht wird, z.B. Modelle für work flows. Diese Prozessmodellierung verwendet hauptsächlich Modelle, die an der Logik orientiert sind, die nur wenige Variable berücksichtigen. Da bin ich auch der festen Überzeugung, dass, wenn man genau schaut, für die Analyse dieser Prozessdaten, statistische Modellierungen von Longitudinaldaten oder Markovprozesse vielfach besser geeignet sind, weil sie eine inhaltliche Komponente einbeziehen können. Ein Regressionsmodell ist immer ein inhaltliches Modell, es ist nicht nur ein Ablaufmodell oder ein logisches Modell und hat nicht nur eine Interpretation der Korrektheit sondern eine inhaltliche Interpretation. Die Idee ein inhaltliches Modells mit einem Modell für die Variabilität zu kombinieren,

das ist ein genuin statistischer Zugang und das ist wichtig und das interessiert mich auch.

**Matthias Templ:** *Eine letzte persönliche Frage: Du bist philosophisch sehr interessiert, habe ich beim gemeinsamen Zugfahren mitbekommen. Hast du die Vorstellung von einem breiten Bildungsbürgertum, dass ein Professor sich praktisch auch anderweitig ...*

**Wilfried Grossmann:** Das kommt von dieser Vorstellung die am Institut für Statistik vorherrschte. Die Vorstellung dass man nicht eine Spezialwissenschaft hat sondern dass man eine gewisse Generalistenhaltung hat, das finde ich wichtig.

Die Interviewer bedanken sich herzlich bei Gabriele Mack-Niederleitner für die Transkription.

## Affiliation:

Wilfried Grossmann
Faculty of Computer Science
University of Vienna
A-1010 Vienna, Austria
E-mail: wilfried.grossmann@univie.ac.at
URL: http://cs.univie.ac.at/ke-team/infpers/Wilfried_Grossmann/

Werner Müller
Department of Applied Statistics and Econometrics
Johannes Kepler Universität Linz
A-4040 Linz, Austria
E-mail: werner.mueller@jku.at
URL: http://www.jku.at/ifas

Matthias Templ
Vienna University of Technology &
Statistics Austria
A-1040 Vienna, Austria
E-mail: matthias.templ@gmail.com
URL: https://www.statistik.tuwien.ac.at/public/templ