

IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOUR UNTUK PREDIKSI WAKTU KELULUSAN MAHASISWA

Irwan Budiman¹, Dodon Turianto Nugrahadi², Radityo Adi Nugroho³
Universitas Lambung Mangkurat^{1,2,3}
irwan.budiman@unlam.ac.id¹, dodonturianto@unlam.ac.id²
radityo.adi@unlam.ac.id³

ABSTRACT

A Collection of data on academic information system database Higher Education is often not fully utilized, but from that data with data mining techniques can provide previously unknown knowledge. Prediction timely graduation of students can be done using k-Nearest Neighbor (k-NN) which is a method to perform the classification of objects based on the training data that was located closest to the object. Prediction timely graduation of students using the k-NN algorithm is done with the stages of data selection, data transformation, data mining and interpretation. This study uses 154 data training and data testing. The accuracy of predictions based on data testing taken by random is 90%.

Keywords: *K-NN Algorithm; Data Mining; Prediction; Classification.*

ABSTRAK

Koleksi data pada *database* sistem informasi akademik Perguruan Tinggi sering tidak dimanfaatkan secara maksimal, padahal dari data tersebut dengan teknik *data mining* dapat memberikan pengetahuan yang belum diketahui sebelumnya. Prediksi waktu kelulusan mahasiswa dapat dilakukan dengan menggunakan algoritma k-Nearest Neighbor (k-NN) yang merupakan sebuah algoritma untuk melakukan klasifikasi terhadap objek berdasarkan *data training* yang jaraknya paling dekat dengan objek tersebut. Prediksi waktu kelulusan mahasiswa menggunakan algoritma k-NN dilakukan dengan tahapan *data selection*, transformasi data, *data mining* dan *interpretation*. Penelitian ini menggunakan 154 data latih dan 10 data uji. Ketepatan prediksi berdasarkan data uji yang diambil secara acak adalah 90%.

Kata Kunci: Algoritma k-NN; Data Mining; Prediksi; Klasifikasi.

PENDAHULUAN

Sebuah fakta yang tampak pada institusi perguruan tinggi adalah bahwa pesatnya peningkatan jumlah edukasional data, tidak dibarengi dengan pemanfaatan data tersebut secara maksimal untuk peningkatan manajemen (Delavari, 2008). Proses tradisional jika didukung teknik *data mining* dapat membantu untuk menemukan pola, struktur dan pengetahuan yang berharga, yang dapat dimanfaatkan oleh instansi terkait sebagai modal untuk peningkatan manajemen khususnya manajemen pengambilan keputusan.

Prediksi waktu kelulusan mahasiswa merupakan salah satu yang penting bagi manajemen perguruan tinggi, hal ini untuk mengantisipasi penumpukan jumlah mahasiswa yang tidak lulus sesuai waktu normal. Bagi mahasiswa yang diprediksi lulus melebihi waktu normal maka dapat diberikan perlakuan khusus agar mahasiswa-mahasiswa tersebut dapat lulus lebih cepat dari hasil prediksi.

Data Mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Teknik *data mining* merupakan sebuah proses ekstraksi informasi untuk menggali pengetahuan (*knowledge discovery*) dan menemukan pola (*pattern recognition*) pada tumpukan data dalam *database* yang biasanya berskala besar (Larose, 2005).

Classification adalah salah satu fungsi *data mining* yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Han, 2006). Algoritma yang digunakan untuk melakukan fungsi klasifikasi adalah algoritma k-NN. K-NN adalah metode klasifikasi yang menentukan kategori berdasarkan mayoritas kategori pada k-NN dilakukan dengan mencari kelompok k objek dalam *data training* yang paling dekat (mirip) dengan objek pada data baru atau *datatesting* (Wu, 2009).

Sistem Informasi Akademik atau yang dikenal dengan SIA pada perguruan tinggi merupakan kumpulan data akademik dan potensinya belum dimanfaatkan secara maksimal sehingga perlu dikelola untuk mencari informasi baru. Data yang digunakan adalah data Indeks Prestasi (IP) mahasiswa mulai dari semester satu sampai dengan semester empat menggunakan salah satu teknik *data mining* yaitu metode klasifikasi dengan menggunakan algoritma k-NN dengan harapan dapat memprediksi waktu kelulusan mahasiswa.

Adapun data training yang digunakan pada penelitian ini adalah data mahasiswa program studi S1 Ilmu Komputer FMIPA Universitas Lambung Mangkurat mulai angkatan 2006 hingga 2009 yang lulus pada tahun 2011 hingga 2015, yaitu sebanyak 154 mahasiswa. Sedangkan data testing menggunakan 10 data mahasiswa angkatan 2010 yang diambil secara acak.

METODE PENELITIAN

Metode penelitian menggunakan tahapan dari proses *Knowledge Discovery from Database* disingkat KDD (Fayyad, 1996), yaitu:

1. *Data Selection*

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalan informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari *database* operasional.

2. *Pre-processing/Cleaning*

Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses *enrichment*, yaitu proses “memperkaya” data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

3. *Transformation*

Proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam *database*.

4. *Data Mining*

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

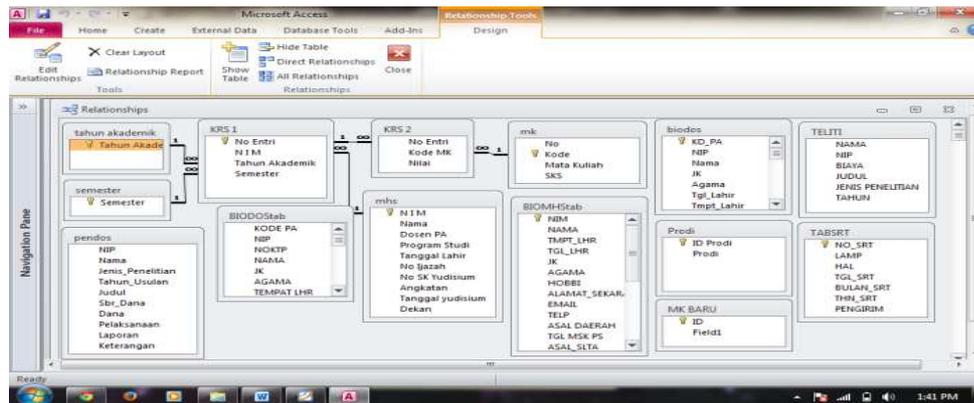
5. *Interpretation/Evaluation*

Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

HASIL DAN PEMBAHASAN

1. *Data Selection*

Tahap menyeleksi data dari *database* yang akan digunakan dalam penelitian. *Database* yang digunakan dalam penelitian adalah *database* Sistem Informasi Akademik. Berikut merupakan relasi tabel yang ada pada *database* Sistem Informasi Akademik sebelum dilakukan penyeleksian data.



Gambar 1. Database SIA

2. Data Cleaning

Semua data sudah relevan, tidak *missing value* dan tidak *redundant*, dan sesuai dengan atribut yang ditentukan. Hal ini merupakan syarat awal dalam melakukan data mining. Dalam *database* ini, terdapat dataset alumni yang berjumlah 154 data.

3. Data Transformation

Tahap membuat dataset baru untuk diproses dengan teknik *data mining*. Dataset yang dibutuhkan mempunyai atribut antara lain NIM mahasiswa, IP semester 1, IP semester 2, IP semester 3, IP semester 4 dan lama kuliah dalam tahun. Dataset yang dihasilkan pada tahap ini adalah seperti pada tabel berikut:

Tabel 1 Dataset prediksi lulus tepat waktu

Mhs	IP1	IP2	IP3	IP4	Lama
1	2,39	2,61	2,74	3	5
2	2,77	2,42	2,86	2,21	6
...
154	3	3	3,79	3,19	5

4. Data Mining

Tahap *data mining* merupakan tahap inti dari penambangan data itu sendiri, dimana penggunaan algoritma k-NN ada pada tahap ini. *Query* yang terakhir pada tahap tranformasi data digunakan sebagai *data training* yang akan dieksekusi menggunakan algoritma k-NN.

Tahapan *query* berikutnya merupakan proses klasifikasi menggunakan algoritma k-NN dimana diperlukan data masukan berupa IP semester 1, 2, 3 dan 4 untuk mencari posisi klasifikasi pada data uji tersebut. Adapun nilai k pada klasifikasi adalah 5. Dan *query*-nya adalah sebagai berikut:

```
SELECT Table1.[NIM], Table1>Nama, Table1.IP1, Table1.IP2, Table1.IP3,
Table1.IP4, Table1.THN,IIf([thn]<=5,"<=5 TAHUN", "> 5 TAHUN") AS
keterangan, [A] AS Expr1, [B] AS Expr2, [C] AS Expr3, [D] AS Expr4,
sqr(((IP1)-[A])^2+((ip2)-[b])^2+((ip3)-[c])^2+((ip4)-[d])^2) AS Expr5
FROM Table1 INNER JOIN Q_ ON Table1.[NIM] = Q_.[NIM]ORDER BY
((IP1)-[A])^2+((ip2)-[b])^2+((ip3)-[c])^2+((ip4)-[d])^2;
```

Adapun *data testing* yang sebanyak 10 data digunakan untuk menguji tingkat ketepatan prediksinya. Berikut ini adalah 10 data IP mahasiswa dari semester 1 sampai dengan semester 4 dan lama kuliah yang sudah diketahui dalam satuan tahun sebagai data uji.

Tabel 2 *Data Testing* 10 Mahasiswa

No	IP smt 1	IP smt 2	IP smt 3	IP smt 4	Lama (tahun)
1	2,61	1,81	2,80	3,46	5
2	2,64	2,62	2,85	3,18	4
3	1,84	1,80	2,67	3,07	5,5
4	2,41	1,56	3,13	2,94	4,5
5	1,95	0,78	1,93	2,47	5,5
6	2,98	2,33	2,94	2,81	5
7	3,02	2,19	2,73	3,52	5
8	1,82	1,60	2,58	3,00	5,5
9	1,02	0,75	1,58	1,70	5,5
10	2,41	2,22	2,94	3,27	4,5

Langkah-langkah algoritma k-NN pada data uji mahasiswa pertama dengan IP semester 1=2,61; IP semester 2=1,81; IP semester 3=2,80; IP semester 4=3,46 dan lama kuliah=5 tahun adalah sebagai berikut:

1. Menentukan nilai k, dalam hal ini k yang digunakan = 5, karena 5 merupakan k yang paling optimal (Banjarsari, 2015).
2. Menghitung jarak *euclidean* data ke-1 sampai dengan data ke-n (154) pada data training, sehingga didapatkan jarak sebagai berikut:

Tabel 3 Hasil perhitungan rumus *euclidean*

No	Rumus	Hasil
1	$(2,39-2,61)^2 + (2,61-1,81)^2 + (2,74-2,81)^2 + (3,00-3,46)^2$	0,9036
2	$(2,77-2,61)^2 + (2,42-1,81)^2 + (2,86-2,81)^2 + (2,21-3,46)^2$	1,9638
...
154	$(3,00-2,61)^2 + (3,00-1,81)^2 + (3,79-2,81)^2 + (3,19-3,46)^2$	2,6212

3. Mengurutkan data jarak (dari kecil ke besar) dan mengambil sebanyak k (5) data teratas, sehingga didapatkan 5 deret data sebagai berikut:

Tabel 4 Hasil Uji pada mahasiswa ke-1

No	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	2,64	1,76	3,1	3,31	5	<=5	0,1159
2	2,79	1,86	3,12	3,48	4	<=5	0,1377
3	2,71	1,79	2,75	3,04	5	<=5	0,1893
4	2,71	2,17	3,06	3,5	4	<=5	0,2088
5	2,91	2,16	2,97	3,4	5	<=5	0,245

4. Pilih data kondisi terbanyak. Dalam hal ini terlihat pada data mayoritas atau 5 dari 5 data menunjukkan masa kuliahnya ≤ 5 tahun.
5. Bandingkan dengan data faktanya. Pada data uji mahasiswa pertama dengan IP semester 1=2,61; IP semester 2=1,81; IP semester 3=2,80; IP semester 4=3,46 dan lama kuliah=5 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola (Hasil diurutkan dari yang terkecil hingga yang terbesar / *ascending*). Mayoritas 5/5 masa kuliahnya ≤ 5 tahun, dimana hal ini tepat dan sesuai dengan keadaan mahasiswa pertama yang waktu kuliahnya 5 tahun.

Langkah-langkah tersebut diatas diulangi lagi pada data uji ke-2 hingga ke-10. Pada data uji mahasiswa kedua dengan IP semester 1=2,64; IP semester 2=2,62; IP semester 3=2,85; IP semester 4=3,18 dan lama kuliah=4 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola. Mayoritas 3/5 masa kuliahnya ≤ 5 tahun, dimana hal ini tepat dan sesuai dengan keadaan mahasiswa ke-2 yang waktu kuliah kurang dari 5 tahun.

Tabel 5 Hasil Uji Pada Mahasiswa Ke-2

No	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	2,43	2,67	3	3,19	5	≤ 5	0,0692
2	2,66	2,34	2,92	3,17	6	> 5	0,0838
3	2,43	2,63	2,64	3,05	5	≤ 5	0,1052
4	2,71	2,31	2,89	3,12	7	> 5	0,1062
5	2,71	2,79	2,95	3,43	5	≤ 5	0,1063

Pada data uji mahasiswa ke-3 dengan IP semester 1=1,84; IP semester 2=1,80; IP semester 3=2,67; IP semester 4=3,07 dan lama kuliah=5,5 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola. Mayoritas 3/5 masa kuliahnya > 5 tahun, dimana hal ini tepat dan sesuai dengan keadaan mahasiswa ke-3 yang waktu kuliah lebih dari 5 tahun.

Tabel 6 Hasil Uji Pada Mahasiswa Ke-3

NO	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	2	1,78	2,97	3,02	6	> 5	0,118
2	2,14	2,03	2,66	3,05	7	> 5	0,143
3	2,21	2,22	2,71	3,21	5	≤ 5	0,334
4	2,21	2,22	2,82	3,19	6	> 5	0,350
5	2,43	1,47	2,71	2,86	5	≤ 5	0,502

Pada data uji mahasiswa ke-4 dengan IP semester 1=2,41; IP semester 2=1,56; IP semester 3=3,13; IP semester 4=2,94 dan lama kuliah=4,5 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola. Mayoritas 4/5 masa kuliahnya ≤ 5 tahun, dimana hal ini tepat dan sesuai dengan keadaan mahasiswa ke-4 yang waktu kuliah kurang dari 5 tahun.

Tabel 7 Hasil Uji Pada Mahasiswa Ke-4

NO	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	2,57	1,95	3,16	2,92	5	<=5	0,179
2	2,43	1,47	2,71	2,86	5	<=5	0,1913
3	2,36	1,89	3,44	3,08	5	<=5	0,2271
4	2,64	1,76	3,1	3,31	5	<=5	0,2307
5	2,57	1,45	2,96	2,52	6	>5	0,243

Pada data uji mahasiswa ke-5 dengan IP semester 1=1,95; IP semester 2=0,78; IP semester 3=1,93; IP semester 4=2,47 dan lama kuliah=5,5 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola. Mayoritas 4/5 masa kuliahnya >5 tahun, dimana hal ini tepat dan sesuai dengan keadaan mahasiswa ke-5 yang waktu kuliah lebih dari 5 tahun.

Tabel 8 Hasil Uji Pada Mahasiswa Ke-5

NO	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	2,32	1,55	2,33	2,59	7	>5	0,9042
2	2,5	1,57	2,16	2,61	6	>5	0,9991
3	2,34	1,61	1,89	2	7	>5	1,0635
4	2,36	1,64	2,27	2,76	5	<=5	1,1074
5	2,07	1,78	2,06	2,78	6	>5	1,1274

Pada data uji mahasiswa ke-6 dengan IP semester 1=2,98; IP semester 2=2,33; IP semester 3=2,94; IP semester 4=2,82 dan lama kuliah=5 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola. Mayoritas 5/5 masa kuliahnya <=5 tahun, dimana hal ini tepat dan sesuai dengan keadaan mahasiswa ke-6 yang waktu kuliah kurang dari 5 tahun.

Tabel 9 Hasil Uji Pada Mahasiswa Ke-6

NO	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	3,07	2,17	3	2,8	5	<=5	0,0374
2	2,86	2,5	2,81	2,93	4	<=5	0,0746
3	3,07	2,27	3	3,11	5	<=5	0,1053
4	3	2,52	3,19	2,96	5	<=5	0,1215
5	2,64	2,19	3	2,79	5	<=5	0,1392

Pada data uji mahasiswa ke-7 dengan IP semester 1=3,02; IP semester 2=2,19; IP semester 3=2,73; IP semester 4=3,52 dan lama kuliah=5 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola. Mayoritas 5/5 masa kuliahnya <=5 tahun, dimana hal ini tepat dan sesuai dengan keadaan mahasiswa ke-7 yang waktu kuliah kurang dari 5 tahun.

Tabel 10 Hasil Uji Pada Mahasiswa Ke-7

NO	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	2,91	2,16	2,97	3,4	5	<=5	0,085
2	2,75	2,24	3	3,35	5	<=5	0,1772
3	2,91	2,47	3,03	3,41	5	<=5	0,1926
4	2,71	2,17	3,06	3,5	4	<=5	0,2058
5	3	2,27	3,17	3,3	5	<=5	0,2488

Pada data uji mahasiswa ke-8 dengan IP semester 1=1,82; IP semester 2=1,60; IP semester 3=2,58; IP semester 4=3,00 dan lama kuliah=5,5 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola. Mayoritas 3/5 masa kuliahnya >5 tahun, dimana hal ini tepat dan sesuai dengan keadaan mahasiswa ke-8 yang waktu kuliah lebih dari 5 tahun.

Tabel 11 Hasil Uji Pada Mahasiswa Ke-8

NO	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	2	1,78	2,97	3,02	6	>5	0,2173
2	2,14	2,03	2,66	3,05	7	>5	0,2962
3	2,07	1,78	2,06	2,78	6	>5	0,4137
4	2,43	1,47	2,71	2,86	5	<=5	0,4255
5	2,36	1,64	2,27	2,76	5	<=5	0,4469

Pada data uji mahasiswa ke-9 dengan IP semester 1=1,02; IP semester 2=0,75; IP semester 3=1,58; IP semester 4=1,70 dan lama kuliah=5,5 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola. Mayoritas 5/5 masa kuliahnya >5 tahun, dimana hal ini tepat dan sesuai dengan keadaan mahasiswa ke-9 yang waktu kuliah lebih dari 5 tahun.

Tabel 12 Hasil Uji Pada Mahasiswa Ke-9

NO	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	2,34	1,61	1,89	2	7	>5	2,6681
2	2	1,54	2,66	1,85	6	>5	2,7734
3	2,43	1,39	2,39	1,85	7	>5	3,0763
4	2,07	1,78	2,06	2,78	6	>5	3,5602
5	2,32	1,55	2,33	2,59	7	>5	3,6846

Pada data uji mahasiswa ke-10 dengan IP semester 1=2,41; IP semester 2=2,22; IP semester 3=2,94; IP semester 4=3,27 dan lama kuliah=4,5 tahun didapatkan pada data training 5 anggota yang jarak data paling mendekati pola. Mayoritas 4/5 masa kuliahnya >5 tahun, dimana hal ini tidak sesuai dengan keadaan mahasiswa ke-10 yang waktu kuliah kurang dari 5 tahun.

Tabel 13 Hasil Uji Pada Mahasiswa Ke-10

NO	IP1	IP2	IP3	IP4	THN	KET	Jarak
1	2,21	2,22	2,82	3,19	6	>5	0,0608
2	2,66	2,34	2,92	3,17	6	>5	0,0873

3	2,21	2,22	2,71	3,21	5	<=5	0,0965
4	2,21	2,47	3	3,21	7	>5	0,1097
5	2,71	2,31	2,89	3,12	7	>5	0,1231

5. Evaluation

Setelah semua data training diproses menggunakan algoritma k-NN dengan nilai k=5 serta jumlah data uji sebesar 10 (tabel 2), maka dihasilkan nilai klasifikasi dan ketepatannya sebesar 90% atau 9 dari 10 data uji, seperti yang ditunjukkan pada tabel berikut ini:

Tabel 14 Prosetase Kesesuaian Prediksi

No MHS	Lama Kuliah (Dalam Tahun)	Prediksi k-NN	Hasil
1	5	<=5	Sesuai
2	4	<=5	Sesuai
3	5,5	>5	Sesuai
4	4,5	<=5	Sesuai
5	5,5	>5	Sesuai
6	5	<=5	Sesuai
7	5	<=5	Sesuai
8	5,5	>5	Sesuai
9	5,5	>5	Sesuai
10	4,5	>5	Tidak Sesuai

KESIMPULAN

Prediksi kelulusan tepat waktu mahasiswa menggunakan algoritma k-NN dilakukan dengan tahapan *data selection*, transformasi data, *data mining* dan *interpretation*. Ketepatan prediksi berdasarkan data uji yang diambil secara acak pada data mahasiswa adalah 90%.

DAFTAR PUSTAKA

Banjarsari, Mutiara A. 2015. Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Berdasarkan IP Sampai Dengan Semester 4. Jurnal KLIK, ISSN:2406-7857, Vol 2, No 2

Delavari, N. 2008. *Data mining Application in Higher Learning Institutions*. Faculty of Information Technology, Multimedia University : Malaysia.

Fayyad, U. M. 1996. *Advances In Knowledge Discovery and Data mining*. Camberidge. MA:The MIT Press.

Han, J.,& Kamber, M. 2006. *Data mining Concept and Tehniques*. San Fransisco : Morgan Kauffman.

Larose, D. T. 2005. *Discovering Knowledge in Data*. New Jersey : John Willey & Sons, Inc.

Wu X, Kumar V. 2009. *The Top Ten Algorithms in Data mining*. New York: CRC Press.