Databases from the SIGEO Network of Permanent Forest Censuses

Smithsonian Symposium on Sharing and Sustaining Research Data

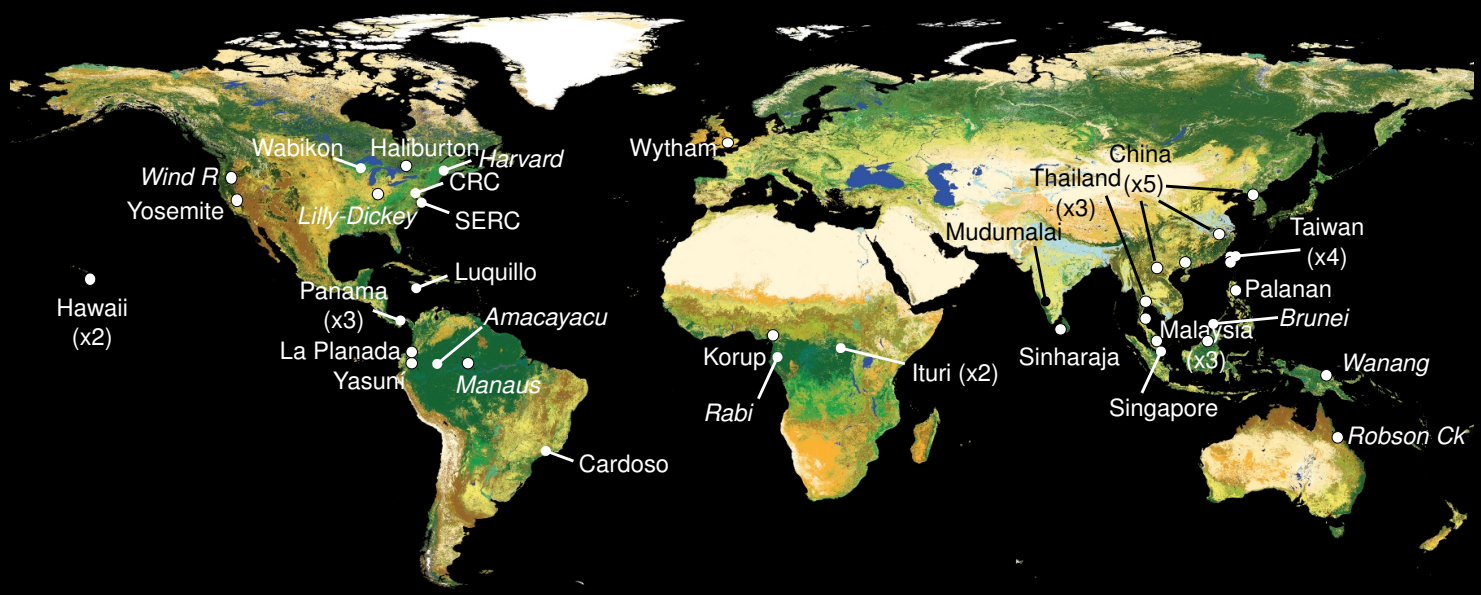
R. Condit, S. Davies, S. Dolins, A. Singh


Smithsonian Tropical Research Institute

SIGEO & CTFS, Smithsonian Tropical Research Institute & Bradley University

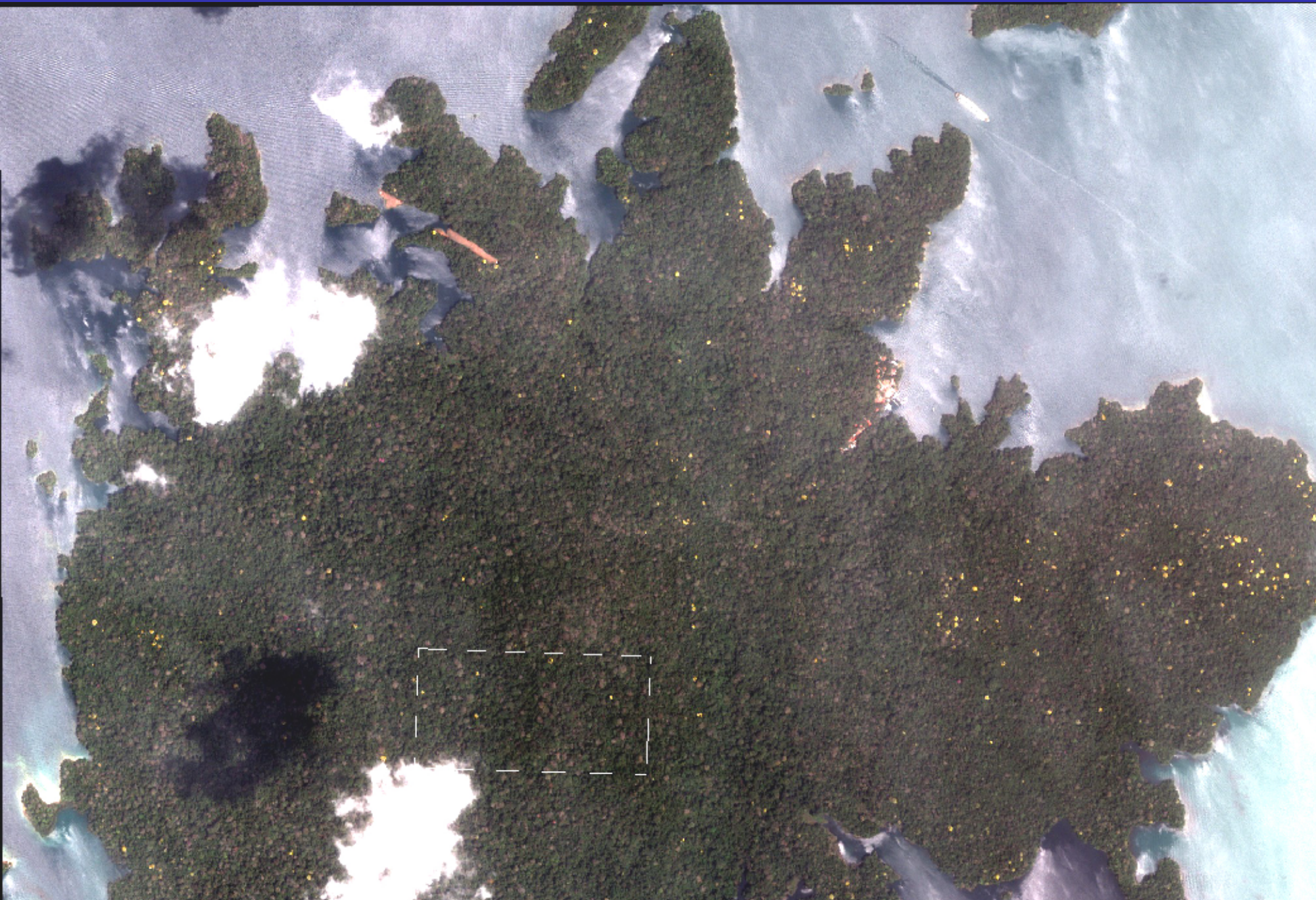
Center for Tropical Forest Science: Smithsonian & Harvard

SIGEO-CTFS-CForBio: Forest censuses following common methods



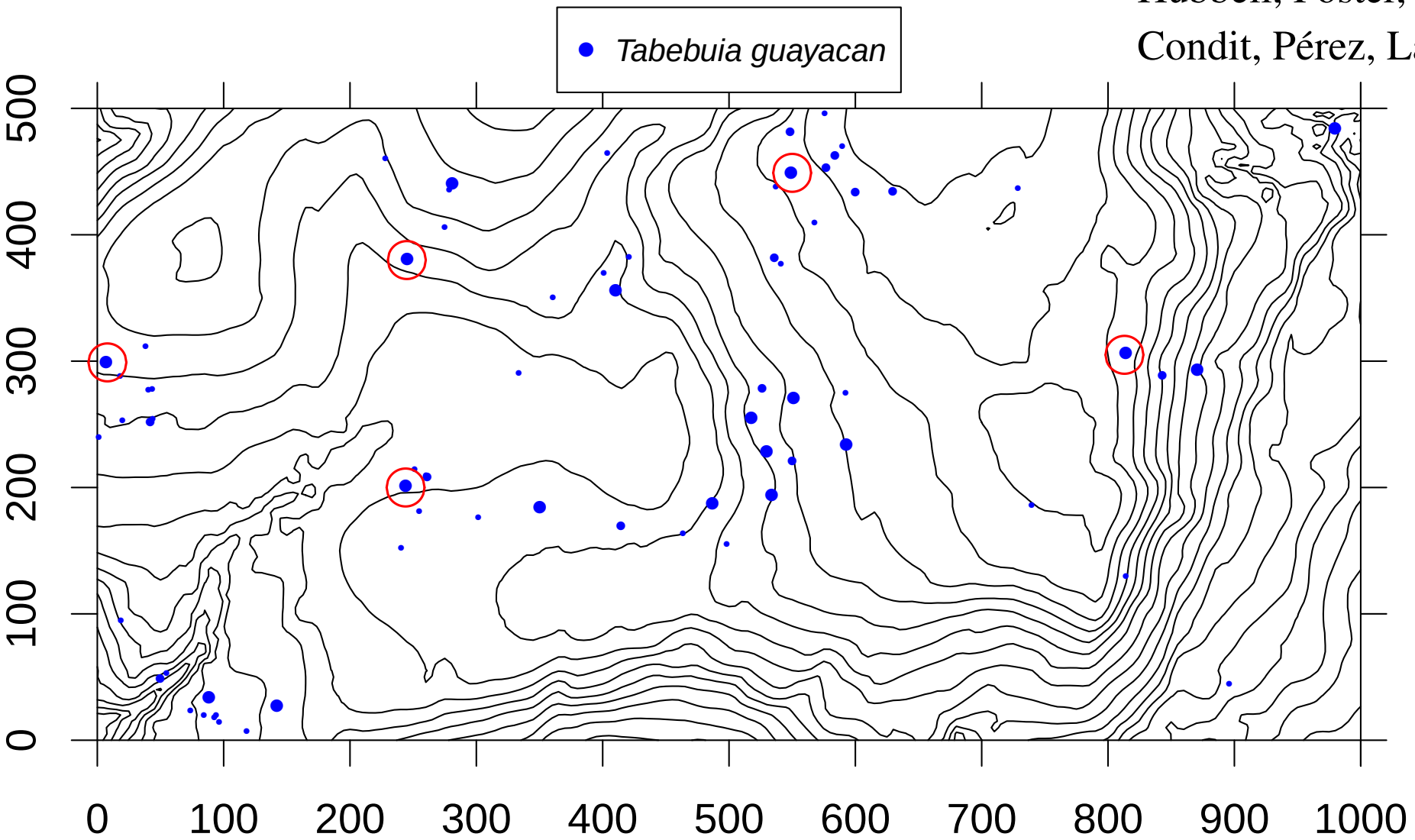
95 completed plots have data in a common database format
-- 3,627,177 trees (ie 3.63×10^6)
-- 8,231 species

SIGEO forest census plots



Barro Colorado census plot

- BCI 50-ha plot
- 1980-2010
- Hubbell, Foster, Condit, Pérez, Lao



Outline

Sample Science, Nature papers from SIGEO network

32 authors completing a comparative example

http://ctfs.arnarb.harvard.edu/Public/pdfs/ConditEtAlScience2006.pdf

A mathematical physicist working with tropical forests

http://ctfs.arnarb.harvard.edu/Public/pdfs/Volkov%20et%20al_Nature_2005.pdf

BCI-Panama forest censuses at STRI

Originated by Hubbell and Foster in 1980

BCI-Panama forest censuses at STRI

Originated by Hubbell and Foster in 1980

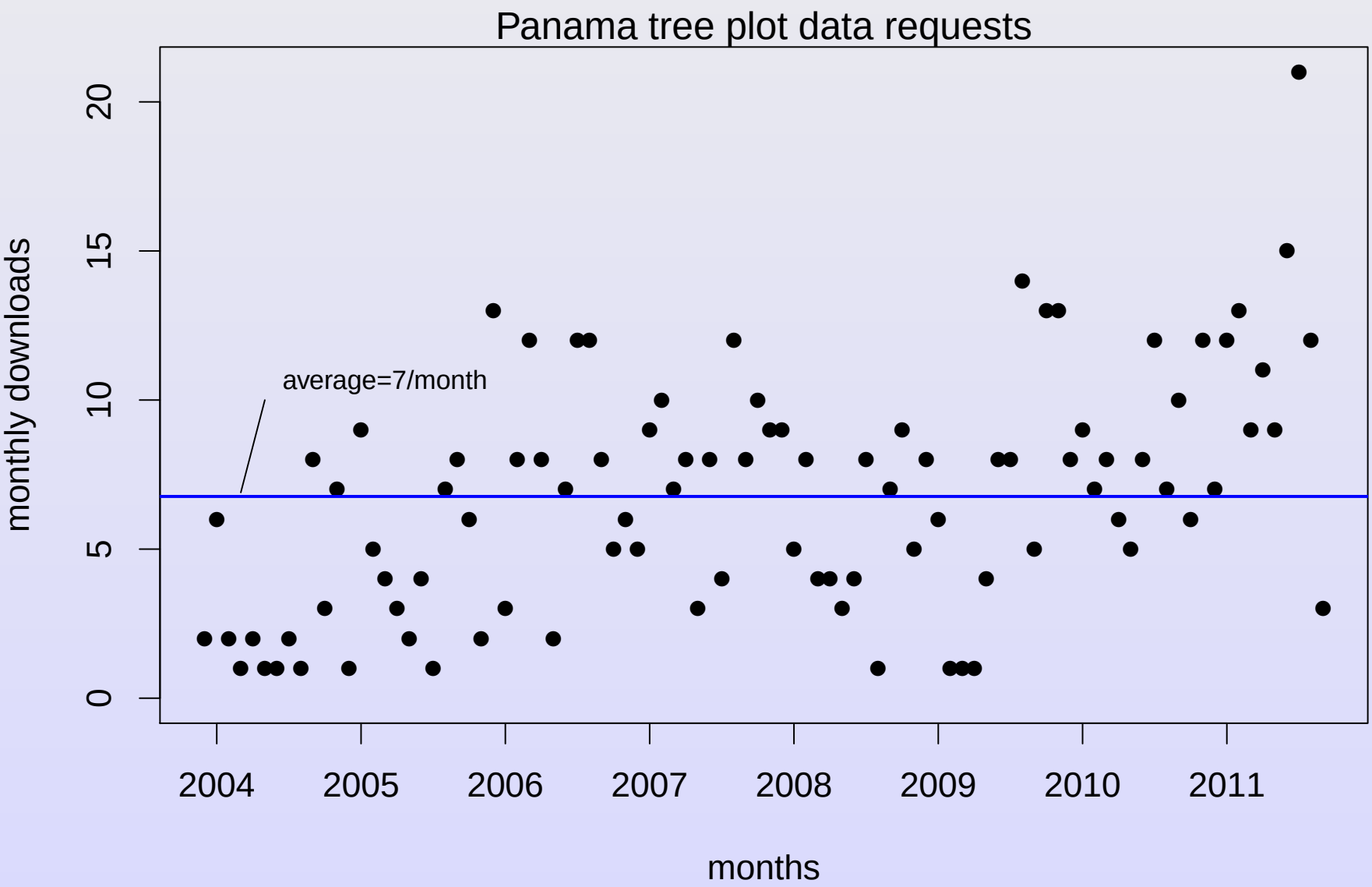
- Seven complete censuses of 50 hectares at BCI
- 64 other plots plus 118 species inventories
- 2.38 million tree measurements
- 30 years of publications (several hundred?)

BCI plot data publicly available

`http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci/`

- First Google hit for either 'BCI 50 ha ...' or 'Panama plot data'
- Online form gathers each request
- 635 requests from 529 individuals since it was announced in *Science* in April 2004

BCI data downloads

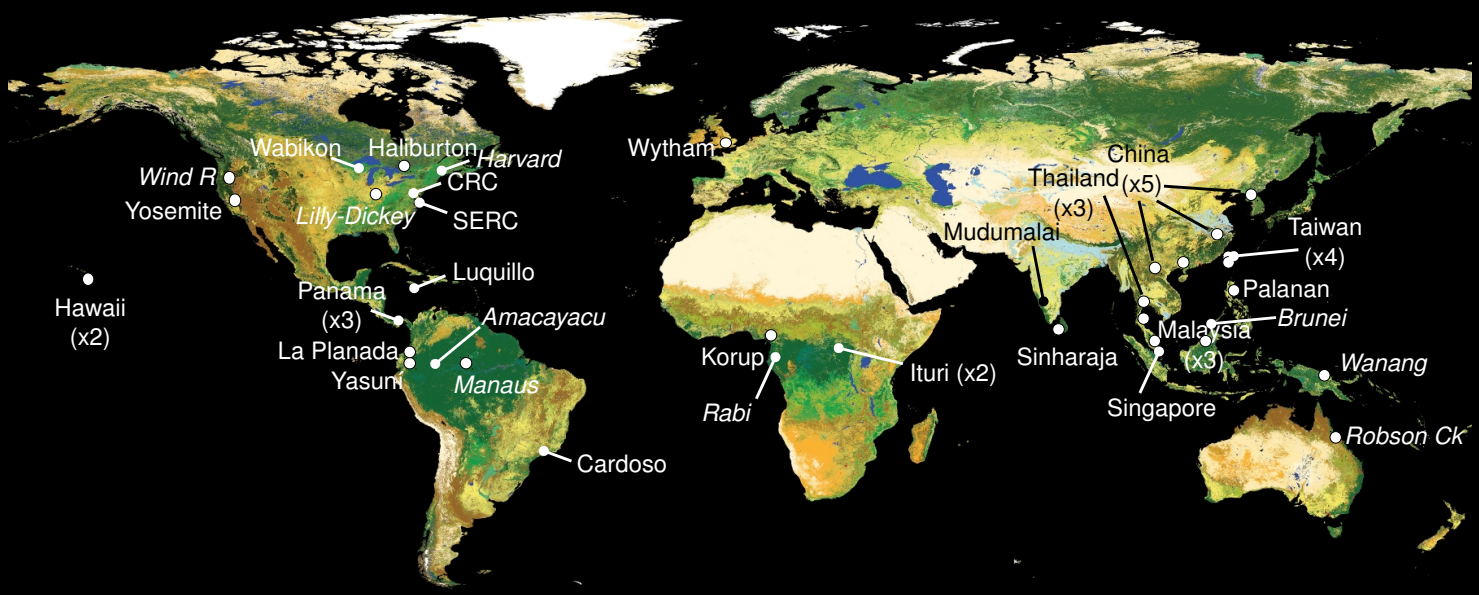


529 BCI data users

- Many for teaching
- Some for data exploration
- Majority are unknown to me and independent of SI
  (I don't have a precise way of tracking this)
- No link for scientific publications back to requests
- No tally of publications

Other SIGEO-CTFS forest plots

SIGEO-CTFS-CForBio: Forest censuses following common methods



95 completed plots have data in a common database format
-- 3,627,177 trees (ie 3.63×10^6)
-- 8,231 species

Other SIGEO-CTFS forest plots

- 46 research projects belong to the network
- 5 of those are within Smithsonian
  SERC, SCBI, Rabi, Mpala, BCI
- 25 of the projects have data managed in standard database system
- Two are publicly available (G. Gilbert, UC Santa Cruz, plus BCI)
- Many others share data on request
- `http://ctfs.arnarb.harvard.edu/Public/Datasets/PlotSummary/AllPlotInfo.php`

Other SIGEO plots

An informal SIGEO-CTFS data access policy

- Data are often made available
- Site PIs are usually co-authors

SIGEO data access: Challenges

Lack of formal procedures for requests

- No clear procedure for data requests
- Depend on informal email contact with PI

Data integrity issues[1]

- Many avoidable errors (eg species misspellings)
- Poorly formatted data difficult to distribute
- Multiple versions of databases in different hands

[1] Data problems solved in our standardized software, but collaborators conflate standardized software with loss of control of data

Reluctance to share data

- Maintaining control of scientific data
- Ensure sole publication access
- Demonstrate ownership

Other SIGEO plots

International collaborators and other[2] cultures' scientific values:

- Do not share scientific goals taken for granted in US/Europe
- Do not value publication in US/European journals
- No interest in data access standards established in US/Europe (for example, standards from journals or granting agencies)

[2]The 5 North American forest plot projects are the most readily shared

Other SIGEO plots

International collaborators and other[2] cultures' scientific values:

- Do not share scientific goals taken for granted in US/Europe
- Do not value publication in US/European journals
- No interest in data access standards established in US/Europe (for example, standards from journals or granting agencies)
- Appreciate assistance with local education
- Seek (country-specific) dissemination

[2]The 5 North American forest plot projects are the most readily shared

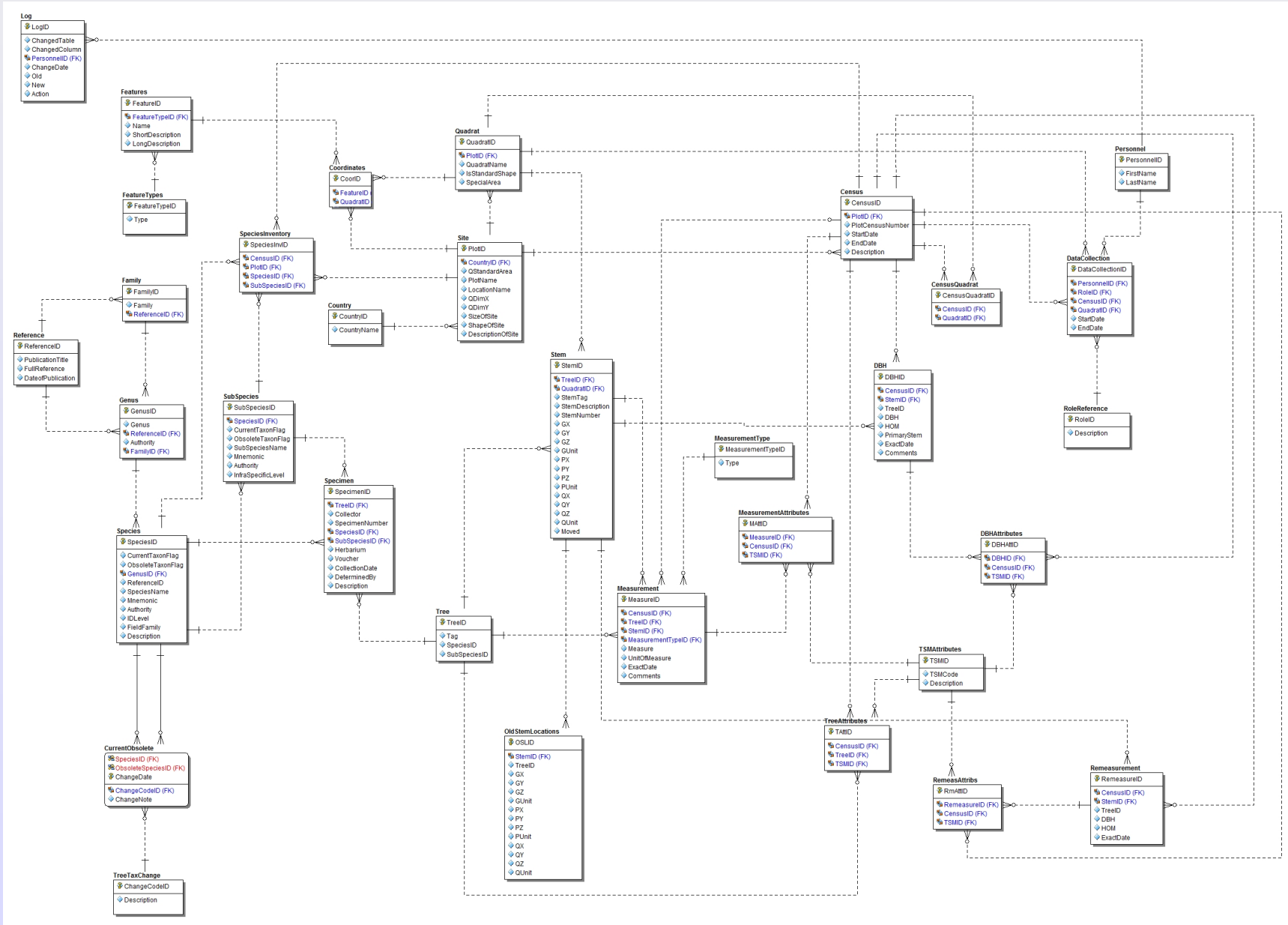
Accomplishments

Rigorous data standards

- Normalized database structure to overcome integrity errors[3]
- Double-data entry software to reduce typos
- Online data to avoid duplicating
- Easy-to-use online reporting system
  `http://ctfs.arnarb.harvard.edu/CTFSReports`
- Analytical software tailored to forest census data
- Broader research agenda in forest ecology with linked databases

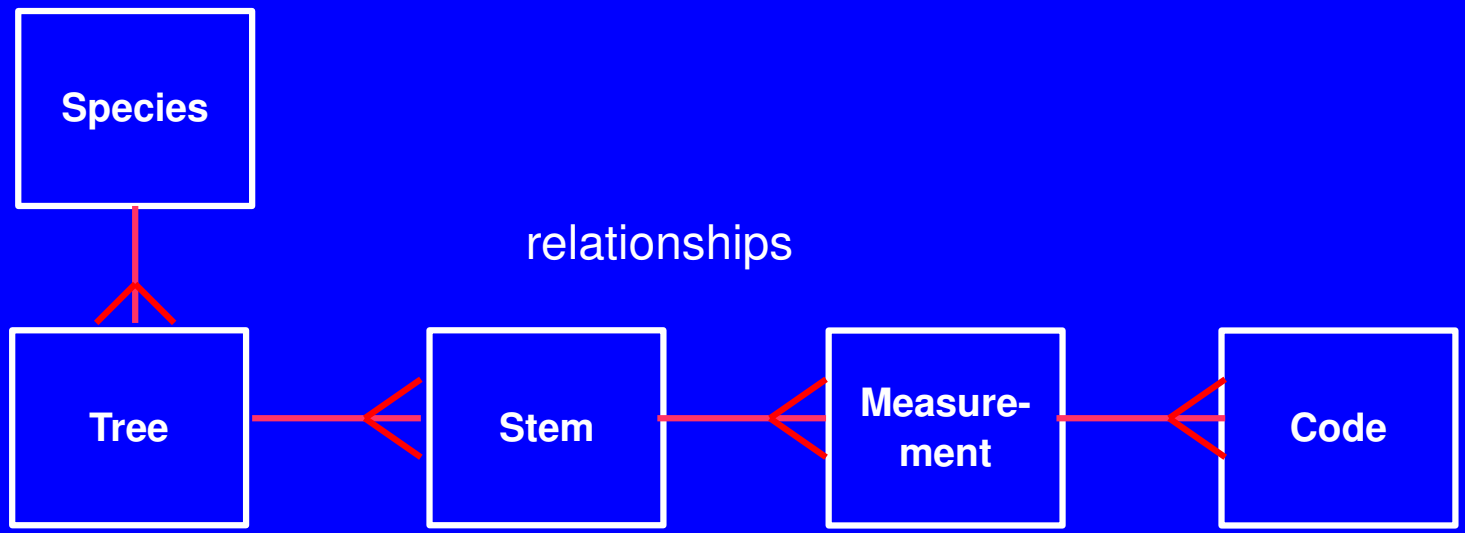
[3] Added complexity of data reduces attractiveness

Database schema



Database schema

Database schema



these are all one to many relationships

Specialized data entry

OLD TREES

P28 - PROYECTO DE LA DINAMICA DEL BOSQUE - CENSO DE 2012

Quadrat: 0201 [Get Form]

Data Collected By: Salomón Aguilar [+]
Data Checked By: Salomón Aguilar [+]
Data Entered By: Salomón Aguilar

Date Collected: April 10 2012
Date Checked: April 10 2012
Date Entered: April 10 2012

Subquadrat Tree Tag	Stem Tag	Species	DBH	Codes	HOM	New DBH	New Codes	Height	Comments
1 , 1									
545801	____	faraoc	45		1.3	____	_____	____	_____
545802	____	pipere	12		1.3	____	_____	____	_____
545803	____	faraoc	58		1.3	____	_____	____	_____
545804	____	faraoc	38		1.3	____	_____	____	_____
545805	____	tropra	216		1.3	____	_____	____	_____
545806	____	hasfl	20		1.3	____	_____	____	_____
545807	____	laciag	67		1.3	____	_____	____	_____
545808	____	faraoc	32		1.3	____	_____	____	_____
545809	____	hirtam	209		1.3	____	_____	____	_____
1 , 2									
545810	____	faraoc	20		1.3	____	_____	____	_____
545811	____	faraoc	56		1.3	____	_____	____	_____
545812	____	virose	43		1.3	____	_____	____	_____
545813	____	faraoc	28		1.3	____	_____	____	_____
545814	____	swarsl	60		1.3	____	_____	____	_____
545815	____	hirtra	27		1.3	____	_____	____	_____
1 , 3									
545816	____	astlat	90		1.3	____	_____	____	_____
545817	____	olmeas	54		1.3	____	_____	____	_____
545818	____	faraoc	32		1.3	____	_____	____	_____

SIGEO-CTFS data access goals

Immediate

Formalize access and restrictions (in place at 2 sites)

- Widely disseminated procedure for data requests
- Online system cataloguing requests (extending BCI system)
- Automatic response explaining site-specific restrictions
- Assuring attribution

SIGEO-CTFS data access goals

Long-term

Release more data

- Publish in data journals
- All data released five years after collection
- Some released immediately with no restrictions

Conclusions

Toward broad data access

- Widely disseminated procedure for data requests
- Online system cataloguing requests (extending BCI system)
- Automatic response explaining site-specific restrictions
- Assuring attribution

- Publish in data journals
- All data released five years after collection
- Some released immediately with no restrictions