

Mixture models for labeling scientific imagery

Michael Turmon, JPL

Co-authors: Kacie Shelton, JPL; Judit Pap, UCLA

To gain a better understanding of the relations between observable solar features and solar irradiance, and to understand the evolution of these features, we wish to segment this imagery into classes conditional on multi-frequency images. We adopt the familiar Markov random field (MRF) models for these feature maps so that spatially coherent labelings are preferred. The MRF models are driven by class-conditional probabilities which we express as mixtures of normal distributions. These distributions may be learned directly from image feature vectors in an unsupervised clustering context. We have found that the resulting clusters do correspond to observable spatial structures. Alternatively, the mixtures may be fitted to expert-segmented images, possibly with constraints expressing prior physical knowledge, or reducing parameter dimension. Use of cross-validated log-likelihood to select the number of components presents no problem, because of the abundance of data. An advantage of the use of mixtures in this setting is that they may be concisely expressed but are also quite flexible, opening up the possibility of a relatively problem-independent solution. In this spirit we developed a portable container for specification of the probabilistic relations between classifications and observables. The textual container lets users (scientists) maintain, annotate, edit, and exchange definite models of their data. These specification documents may be interpreted by inference engines that can sample the model and compute probabilities defined by it. In addition to these general-purpose mechanisms, we also consider special constructs for concise description of temporal and spatial patterns of dependence between random variables.

MIXTURE MODELS FOR LABELING SCIENTIFIC IMAGERY

Michael Turmon

28 July 2001

- A. Introduction
- B. Image Labeling
- C. Data Models
- D. Results

Joint work with Judit Pap of UCLA and
Kacie Shelton of the JPL Data Understanding Group

turmon@aig.jpl.nasa.gov

<http://www-aig.jpl.nasa.gov/home/turmon/>

SCIENTIFIC GOALS

Reliably identify structures in the photosphere

Relate these structures to irradiance changes

Features

- Sunspots
- Faculae: Can be reliably distinguished from sunspots
- Quiet sun: everything else

Cannot now consistently identify network/supergranulation

Methods

Automatic, objective classification using statistical model

Model quantifies the uncertain relation of observables to classes

Model uses spatial information to choose labels

Falsifiable models (Popper 1958) can be checked against
the data they claim to model

General method that extends unchanged to other settings, e.g.

more observables

different number of features

explicit accounting for miscalibration; outliers

inclusion of physical knowledge (like sensor noise)

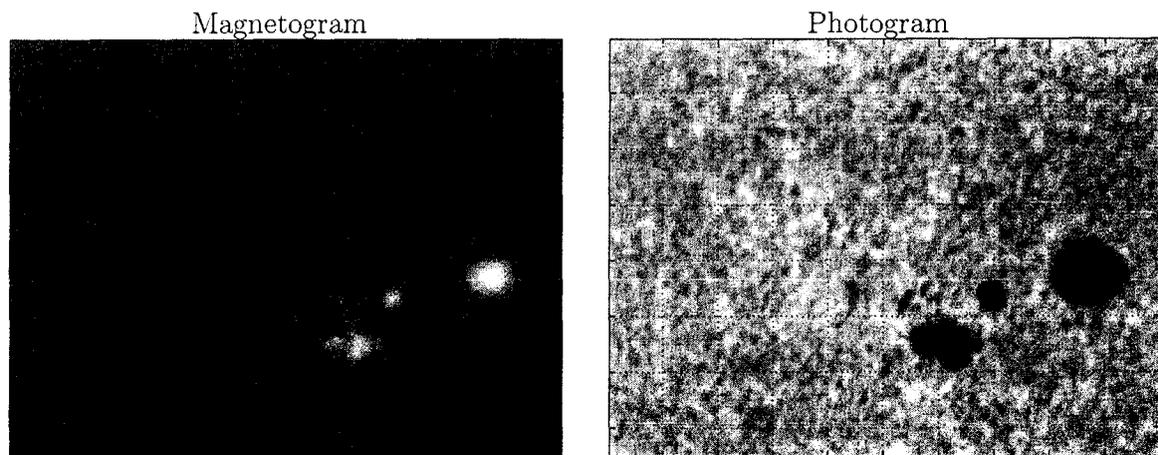
DATA SOURCES

These are irregularly-sampled time series of images

MDI on SoHO

Magnetograms and quasi-photograms having 1Kx1K pixels
Modern & consistent; no night/cloud/atmosphere difficulty
Several/day since May 1996, about 500 images/month

Analyzed May 1996 – Sep 2000; 60 GB across 25 000 images
Using MDI level 1.5 magnetograms and level 2 photograms



Taken by SoHO/MDI in August 1996

Other Sources

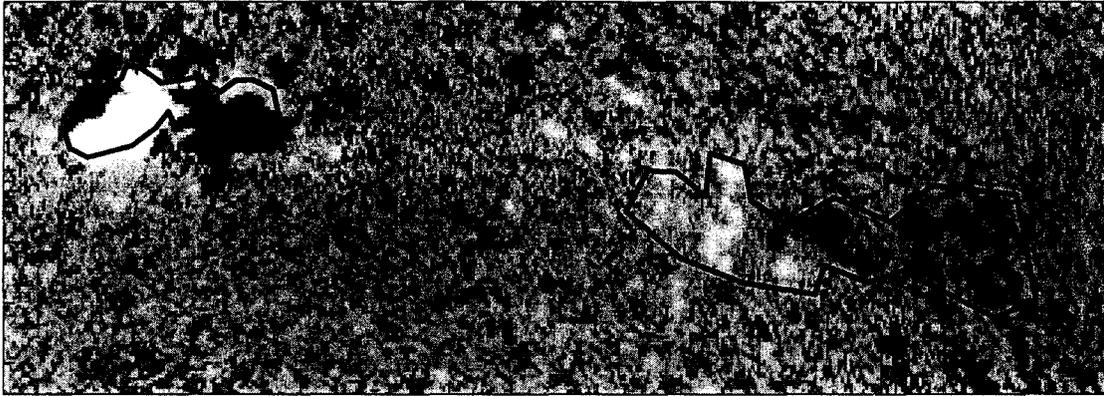
Digitized films from Mt. Wilson Observatory, CA
Magnetic field and intensity images for several decades.

PSPT in Hawaii as cross-calibration

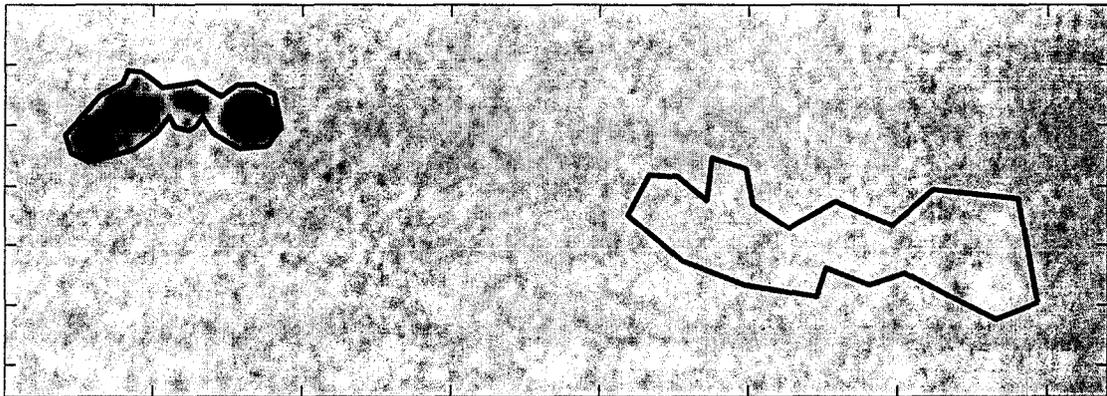
WHY USE BOTH?

17:58 UTC on 7 September 1997

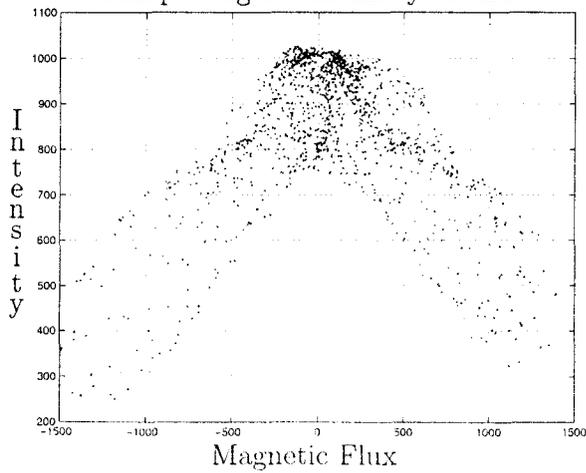
Preprocessed Magnetogram: Detail



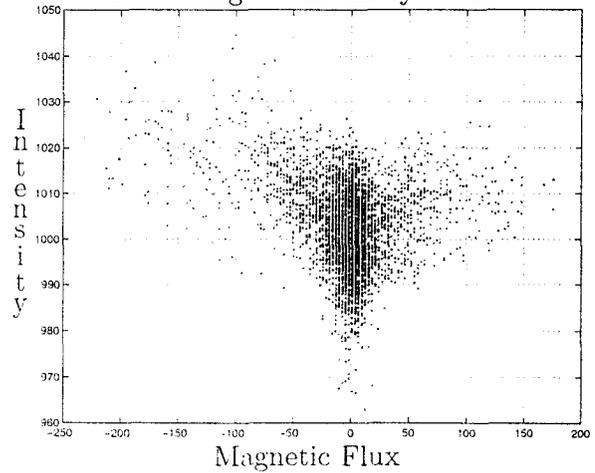
Preprocessed Photogram: Detail



Sunspot region: Intensity vs. flux

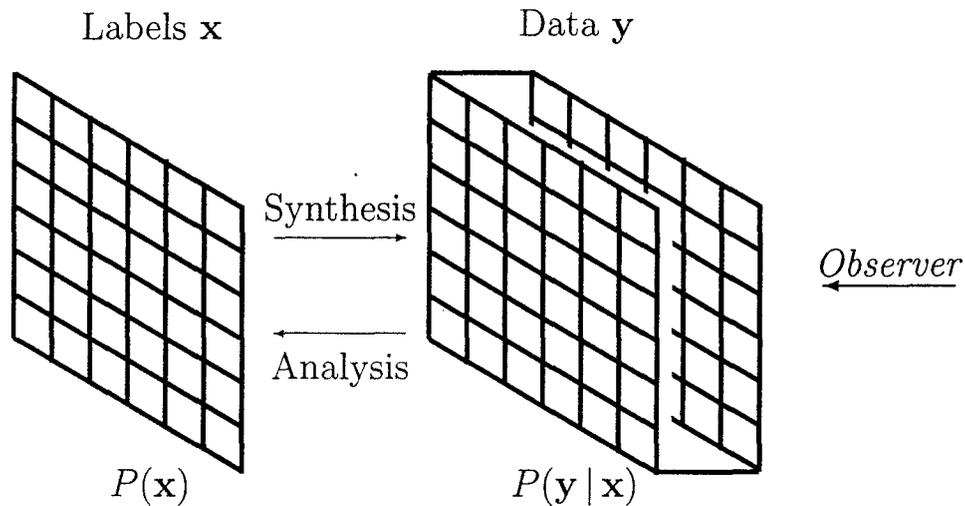


Facula region: Intensity vs. flux



PROBABILISTIC IMAGE MODELS

Quantitatively describe the uncertain relation between observables and labels in an accepted probabilistic framework



At each spatial position, one of K physical processes is dominant.

Observables arise depending on the dominant physical process.

Generation of observables may be viewed as adding uncertainty (noise) to the underlying dominant process.

Goal of analysis is to invert this noisy mapping.

Variables of the Model

Index set \mathcal{N} of spatial coordinates $s = (i, j)$

Unobservable labels $\mathbf{x} = [x_s]_{s \in \mathcal{N}}$ & observables $\mathbf{y} = [\vec{y}_s]_{s \in \mathcal{N}}$

x_s : small integer $1 \dots K$ (i.e. ACR/Fac/QS)

\vec{y}_s : real vector (i.e., the pair (magnetic field, light intensity))

Statistical model given by two distributions $P(\mathbf{x})$ and $P(\mathbf{y} | \mathbf{x})$

MODEL SPECIFICS

Describe the two distributions $P(\mathbf{x})$ and $P(\mathbf{y} | \mathbf{x})$

1: Link to Observables

Make the link via scientist-labeled images and distribution-fitting

Alternatively, can infer automatically from data via clustering

Obtain K distributions, one for each feature class

MDI illustration: posit a (bivariate) Gaussian law

$$P(\vec{y}_s | x_s = k) \sim \text{Normal}(\vec{\mu}_k, \sigma_k^2 I)$$

(QS class, $k = 1$: fits the SoHO/MDI data reasonably well using $\vec{\mu}_1 = [0 \ 1]$ and $\sigma_1 = 0.01$.)

2: Quantifying Spatial Smoothness

Typically $\beta \geq 0$ controls smoothness in the prior

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(-\beta \sum_{s \sim s'} 1(x_s \neq x_{s'})\right)$$

where $s \sim s'$ means: site s close to site s' , e.g. one pixel away

Penalty of β per disagreement of nearby pixels to enforce spatial coherence of labelings

At $\beta = 0$, penalty and spatial constraint vanish

- Objective, automatic inference possible given $\vec{\mu}_k, \sigma_k^2, \beta$

INFERRING THE LABELING

Invert the noisy data the Bayesian way with familiar *maximum a posteriori* (MAP) estimate

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x}} P(\mathbf{x}|\mathbf{y})$$

Bayes rule shows $P(\mathbf{x}|\mathbf{y}) \propto P(\mathbf{y}|\mathbf{x})P(\mathbf{x})$

first factor is the traditional “likelihood function”

second is the prior enforcing spatial coherence

For normal model, algebra reveals the objective function above is

$$\log P(\mathbf{x}|\mathbf{y}) = -\frac{1}{2\sigma^2} \sum_{s \in \mathcal{N}} \|\vec{y}_s - \vec{\mu}_{x_s}\|^2 - \beta \sum_{s \sim s'} 1(x_s \neq x_{s'})$$

Interpretation

First term: fidelity to data (observation close to its mean)

Second term: image smoothness (this couples the pixel labels)

Maximizing $P(\mathbf{x} | \mathbf{y})$ is a numerical problem solved in about 3 min/image on Sun workstation (360MHz).

Algorithm cycles through label pixels, refining it over many sweeps.

Data Models

But, the normal distribution is not adequate for all classes:

it fails standard statistical tests.

...normal model is thus *falsified*.

We must introduce more realistic data models $P(\vec{y} | x)$

ASSESSING VALIDITY

Models for observables

These are directly sensed, allowing direct model checks

Computing $P(\text{data} | \text{model})$ falsifies some models

e.g., normal model for class-conditional distributions
is falsified on these grounds

Models for labels

Known as *ground truth* and is difficult to verify

- If a scientist says it is a sunspot, it may not be a sunspot.
Difficulty with $P_D = P(\text{say sunspot} | \text{is sunspot})$.
- Further: Lack of physical understanding of problem means even experts may be surprised at what is really there.

What is being modeled

Active region discovery formulated as an image segmentation problem

Optimistic to represent photospheric attributes by a discrete class

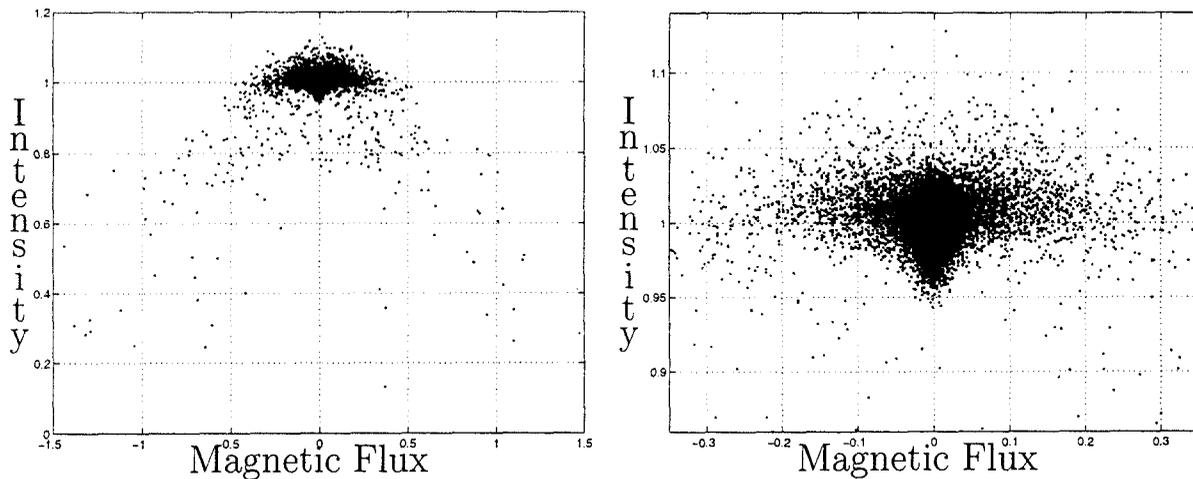
Perhaps represent attributes as a vector $x = [x_1 \ x_2 \ x_3]$, where e.g. x_1 is our subjective belief that it is a sunspot

More expressive schemes possible and perhaps appropriate

MODELING THE OBSERVABLES

Main ingredient: the three *density models*

$$P(\vec{y}_s | Q), \quad P(\vec{y}_s | F), \quad P(\vec{y}_s | \text{ACR})$$



These density models will dominate the image labelings

As strawman, put forward per-class normal distributions

$$P(\vec{y}_s | x_s = k) \sim \text{Normal}(\vec{\mu}_k, \Sigma_k)$$

with $d \times 1$ class means and $d \times d$ covariance matrices.

Simple normal distributions are too simplistic for this data:

strongly multimodal

cannot even transform to normality (e.g., with $|\text{flux}|$)

quiet class, e.g., contains superpositions of effects

(supergranulation is discernable in scatter plots)

Three questions (with answers):

modeling (*normal mixtures*)

fitting (*maximum likelihood via EM algorithm*)

validation (*cross-validation*)

USING NORMAL MIXTURES

Modeling

For sunspot especially, benefit from the flexible density model

$$p(\vec{y}; \theta) = \sum_{g=1}^G \alpha_g N(\vec{y}; \vec{\mu}_g, \Sigma_g)$$

$$\theta = \{(\alpha_1, \vec{\mu}_1, \Sigma_1) \cdots (\alpha_G, \vec{\mu}_G, \Sigma_G)\}$$

Convex combination of gaussian bumps or *normal mixture*

Bump g has weight α_g , centered at $\vec{\mu}_g$ with elliptical contours Σ_g .

Accounts for multimodality and superpositions of effects

A very general family: take G large.

Estimation

Ask scientists to find regions of type $x_s = k$; estimate θ_k for each

Goal: From data $Y = [\vec{y}^1 \cdots \vec{y}^n]$, find a density model $p(\vec{y}; \hat{\theta})$

Method: Determine parameters by maximum-likelihood using Y :

$$\hat{\theta} = \arg \max_{\theta} \log P(Y; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log p(\vec{y}^i; \theta)$$

where $p(\vec{y}; \theta)$ is as above

Performed via EM algorithm, a popular iterative procedure for learning parameters by maximum-likelihood.

...done once and then the model is fixed

Alternatively, can provide cumulative data over classes, and EM will *cluster* vectors \vec{y} into classes

clusters are extracted after the fact: unsupervised learning

MODEL VALIDATION

Overfitting

Find $\hat{\theta}$ from Y , varying number of bumps $G = 1, 2, \dots$

$$\hat{\theta}_G = \arg \max_{\theta} \log P_G(Y; \theta)$$

As G increases, “better” fits $\hat{\theta}_G$ to Y are obtained!

Overfitting phenomenon: too many parameters to fit reliably

Controlling model complexity with cross-validation

Solution: evaluate models on a separate validation data set

Hold aside test data $Z = [\vec{y}^1 \dots \vec{y}^m]$ disjoint from Y

Train $\hat{\theta}_G$ from Y with maximum-likelihood as indicated

Test $\hat{\theta}_G$ on separate data Z

Contrast *test likelihood* $P_G(Z; \hat{\theta}_G)$ with $P_G(Y; \hat{\theta}_G)$: the former is an unbiased estimate of fit of $\hat{\theta}_G$ to true distribution

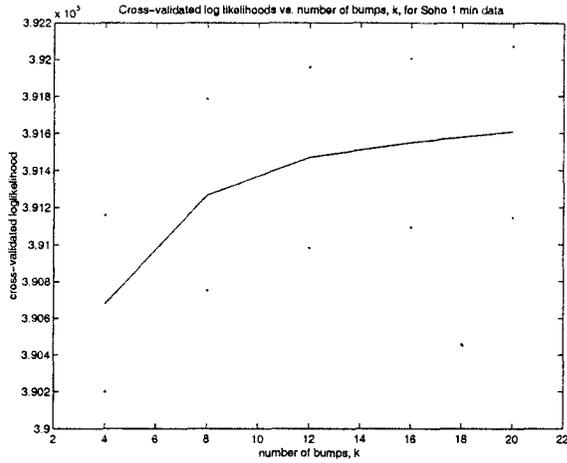
Next, generate more training/test (Y/Z) splits to get more estimates of goodness-of-fit

Average of these goodness-of-fit indicators shows what model complexity the data can support

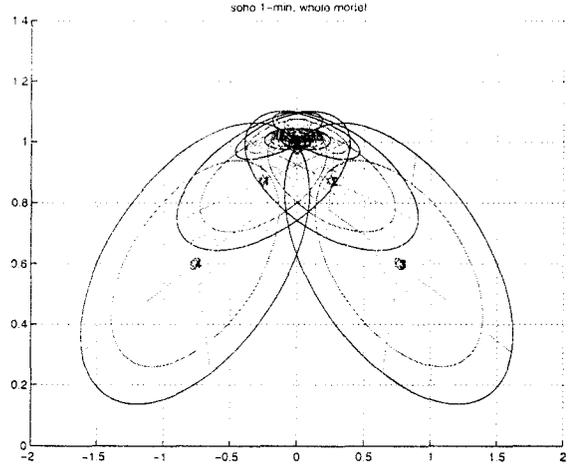
Also, can serve to test robustness of model to changes in data to which it should be largely invariant

MODELS USED

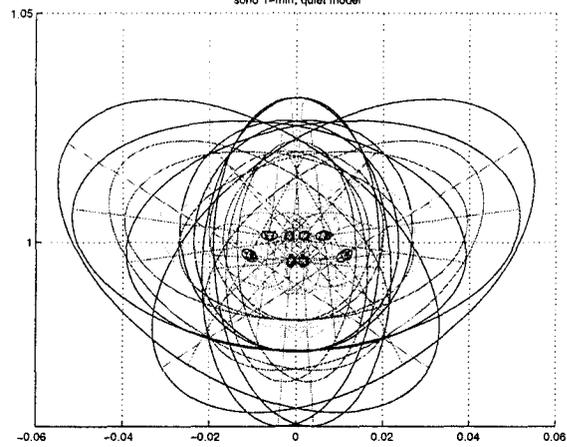
Model Fit, varying Complexity



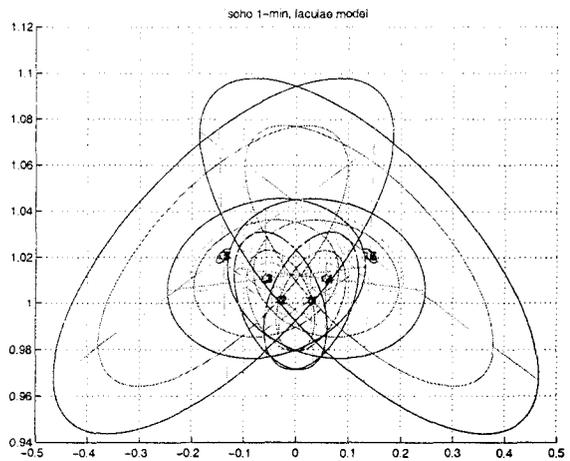
Entire Model



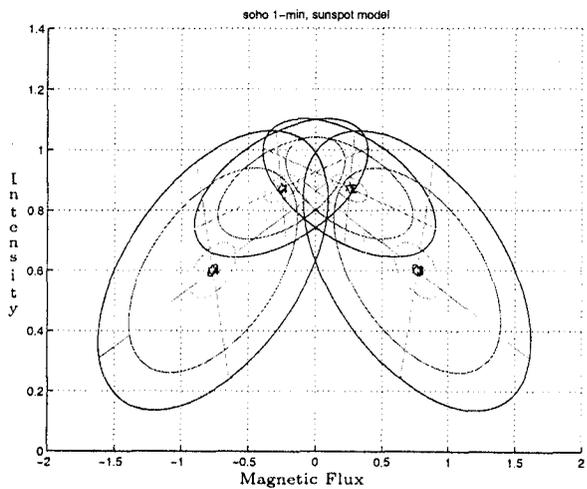
Quiet Sun Model



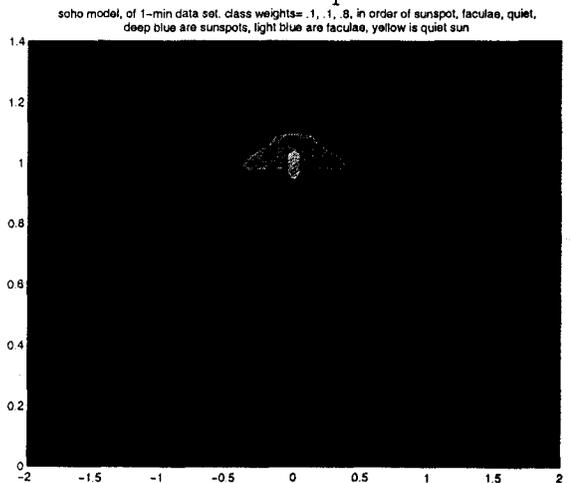
Facula Model



Spot Model

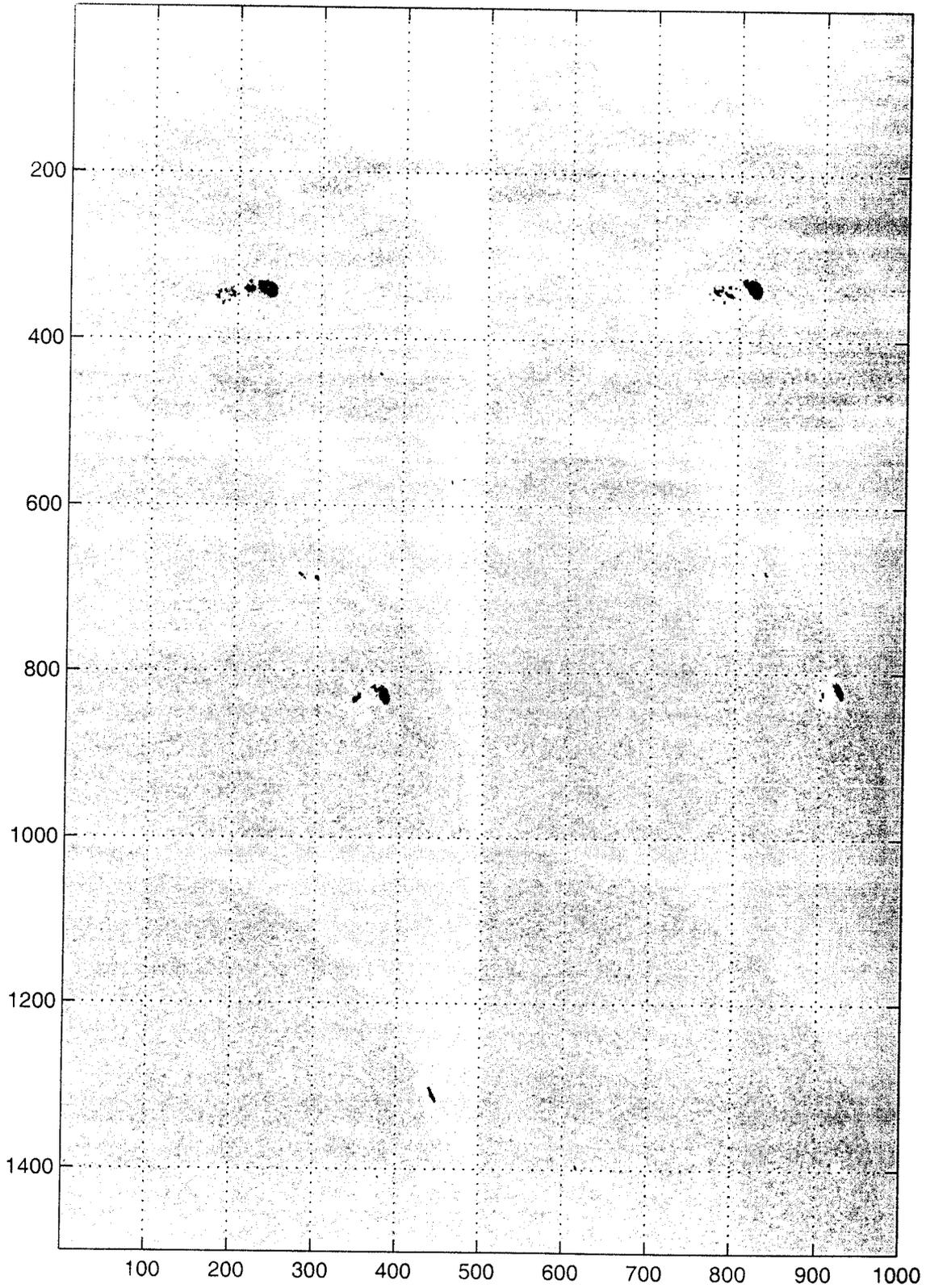


Class Map

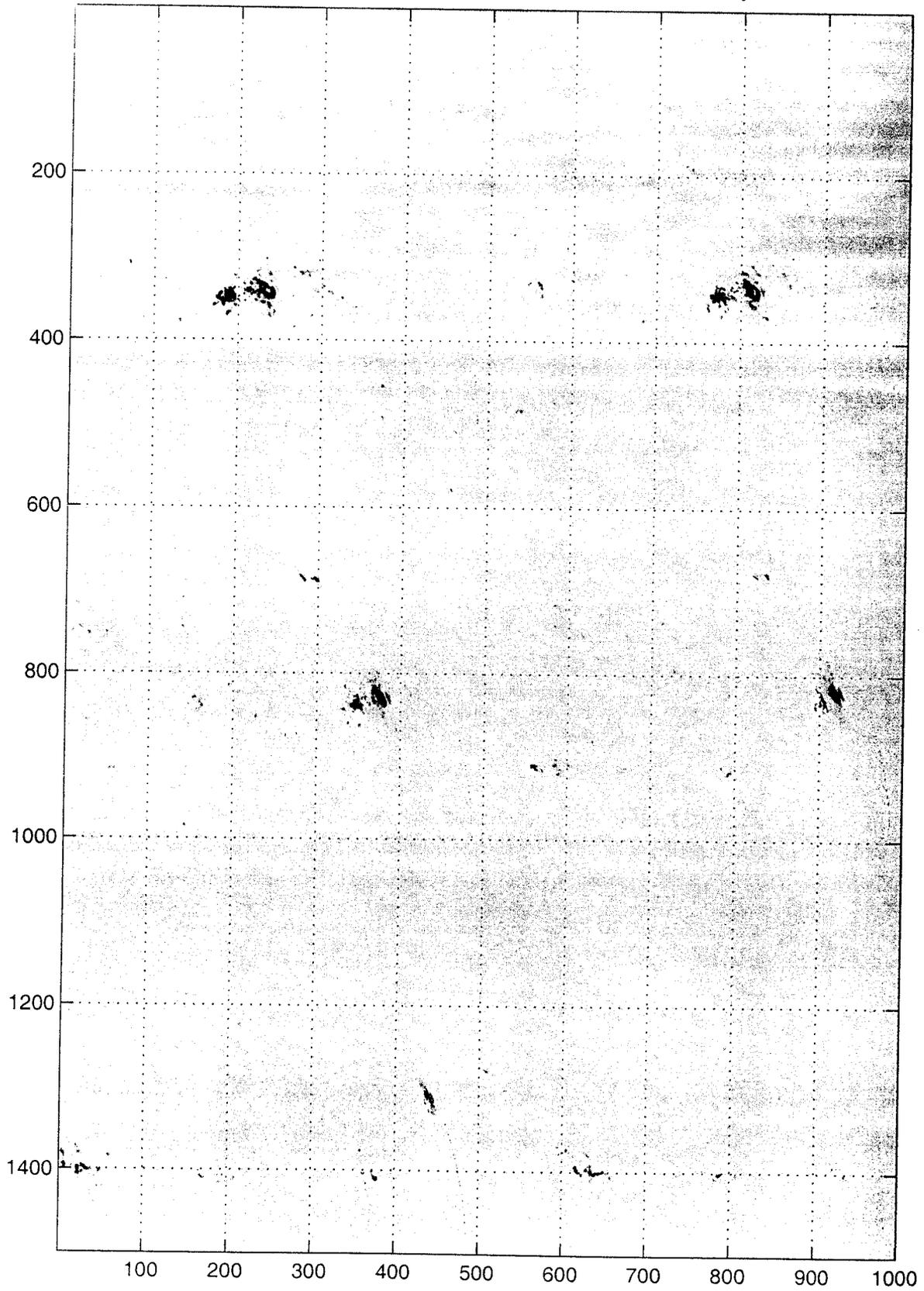


LABELINGS

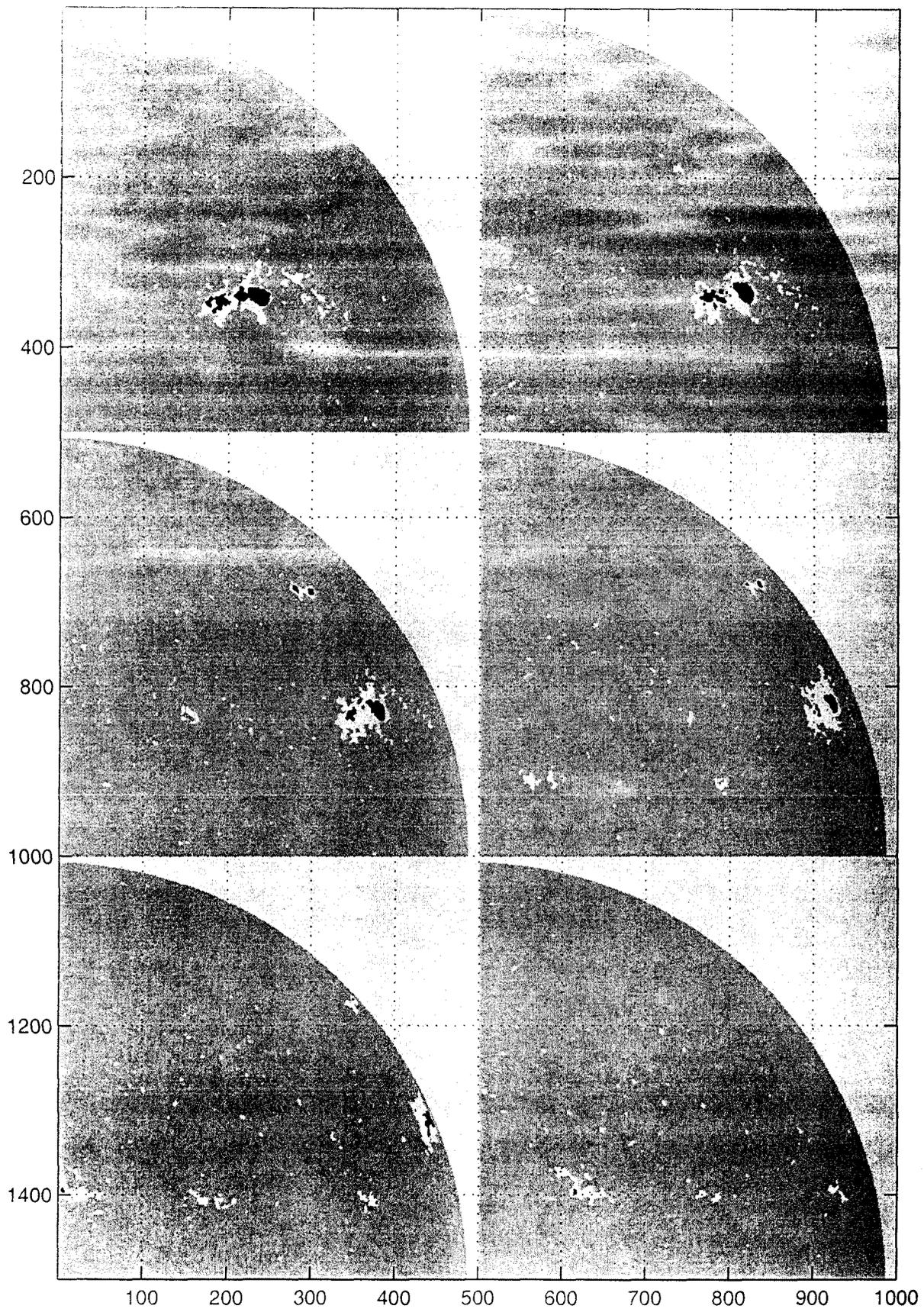
Photogram: 1998/01/15 11:11 UTC + 0,1,2,3,4,5 days



Magnetogram: 1998/01/15 11:11 UTC + 0,1,2,3,4,5 days



Labeling: 1998.01/15 11:11 UTC + 0.1,2,3,4,5 days



CONCLUSIONS

Methods

Framework suited for a variety of labeling problems

Models are fit automatically from flexible family

Leads to falsifiable statistical model for data

Spatially, temporally uniform data is key to accurate labelings

Results

Primary dataset is 60 GB of MDI m- and p-grams over > 4 years

Stable identification of features in time

Excellent agreement between MDI and Mg c/w

Developed a usable inspection, annotation, and automated labeling tool

Futures

More accurate modeling of long-term MDI imagery

Integration, comparison with instruments like PSPT, SDO, MWO

Region tracking for irradiance and other purposes

Object taxonomy by clustering

turmon@aig.jpl.nasa.gov

<http://www-aig.jpl.nasa.gov/home/turmon/>