



# Kartografické modelování

## X – Prediktivní modelování

jaro 2015

**Petr Kubíček**

**kubicek@geogr.muni.cz**

**Laboratory on Geoinformatics and Cartography (LGC)  
Institute of Geography  
Masaryk University  
Czech Republic**



# Podstata prediktivního modeování

- Doposud jsme se zabývali problémem, jak počítač „vidí“ geografická data prostřednictvím popisných (deskriptivních) technik a vytváří z nich oblasti s určitými vlastnostmi.
- Další logický krok je použití „**prediktivních – předpovědních**“ technik k vytvoření **extrapolačních map předvídajících budoucí podmínky**.
- Využití v řadě oblastí:
  - **Predikce kriminality.**
  - Zemědělství – odhady výnosu plodin.
  - Archeologie - lokalizace naleziště.



# Predictive Crime Analysis

- **WHAT?**
- „Predictive policing in the context of place is the use of **historical data** to create a **spatiotemporal forecast** of crime **hot spots**.
- **WHY?**
- that will be the **basis for police resource allocation** decisions with the expectation that having officers at the proposed place and time **will deter or detect criminal activity.**“



# The role of 'place' in crime

Two key considerations (Spencer Chainey)

- Crime has an inherent **geographical quality**
- Crime is **not randomly distributed**

# Crime has an inherent geographical quality

The four dimensions of crime:

- **Legal** (a law must be broken).
- **Victim** (someone or something has to be targeted).
- **Offender** (someone has to do the crime).
- **Spatial** (it has to happen at a place - somewhere, in space and time).

# Crime is not randomly distributed

If crimes were random:

- Equal chance of them happening anywhere at anytime.

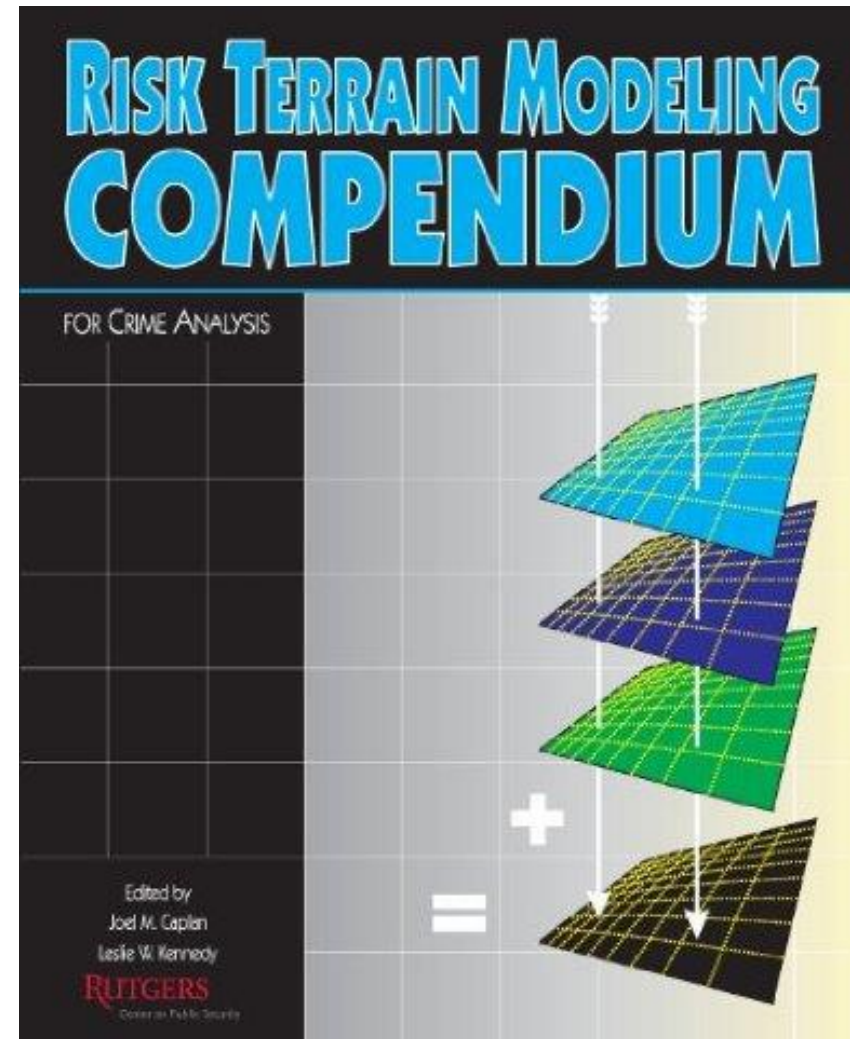
But crime is not randomly distributed

- Concentrated into places of activity
  - Crime hotspots
- Series follow geographic patterns
  - Serious and volume crime

# Risk Terrain Modeling Prediction

- Risk terrain modeling (RTM) is an **approach to risk assessment** in which separate **map layers** representing the influence and intensity of a **crime risk factor** at every place throughout a geography is created in a geographic information system (GIS).
- Map layers are combined to produce a **composite “risk terrain” map** with values that account for all risk factors at every place throughout the geography.
- Available in PDF – ask your lecturer 😊

**Kartografické modelování**



# RTM steps

1. Select an outcome **event** of particular interest
2. Choose a study **area**
3. Choose a time **period**
4. Obtain **base maps** of your study area
5. Identify **aggravating** and **mitigating factors** related to the outcome event
6. **Select** particular **factors** to include in the RTM
7. **Operationalize** the spatial influence of factors to risk map layers
8. **Weight** risk map layers relative to one another
9. **Combine** risk map layers to form a composite map
10. **Finalize** the risk terrain map to **communicate** meaningful and actionable information.



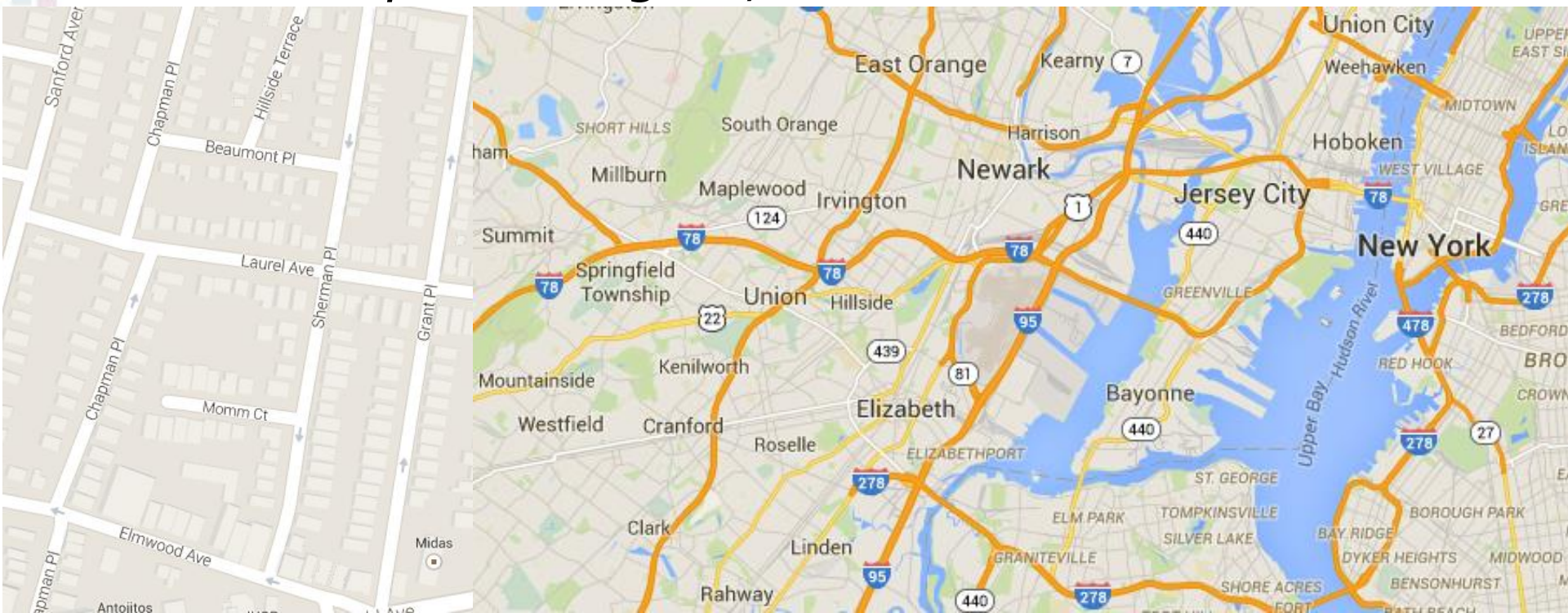
# Step 1 -2

1. Select an outcome **event** of particular interest

*Gun shooting incidents.*

2. Choose a study **area on which risk terrain maps will be created.**

*The Township of Irvington, NJ.*

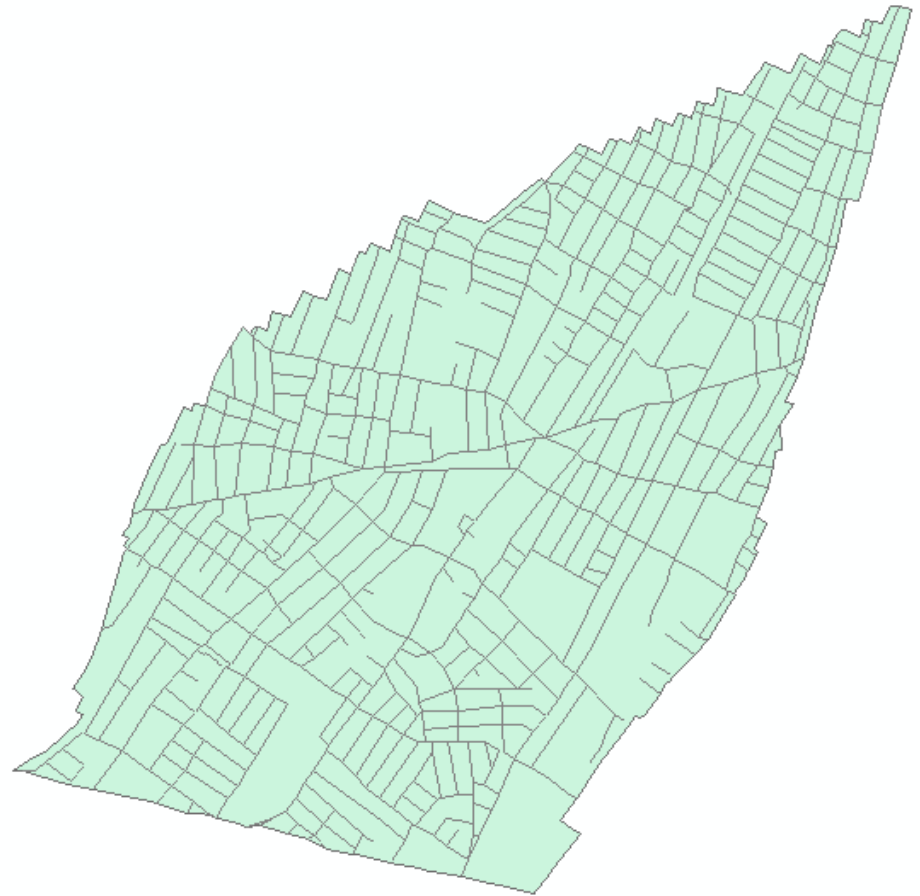


## ***STEP 3: Choose a time period to create risk terrain maps for.***

- Six month time period: January 1 to June 30.
- It is expected that this time period will adequately assess the place-based risk of shootings during the next 6-month time period (July 1 to December 31).
- **Data availability and comparability ?? Is it really justifiable and valid for the Czech Republic?**

## Step 4

- ***STEP 4: Obtain base maps of your study area.***
- Two base maps were obtained from Census 2000 TIGER/Line Shapefiles:
  - 1) Polygon shapefile of the Township and
  - 2) **Street centerline** shapefile for the Township.



## ***STEP 5: Identify aggravating and mitigating risk factors that are related to the outcome event.***

- Three **aggravating factors** were identified based on a ***review of empirical literature***:
  - dwellings of known gang members (**habitual offenders**),
  - locations of **retail business infrastructure** (bars, strip clubs, bus stops, check cashing outlets, pawn shops, fast food restaurants, and liquor stores),
  - locations of **drug arrests** (places, where the police action happened).

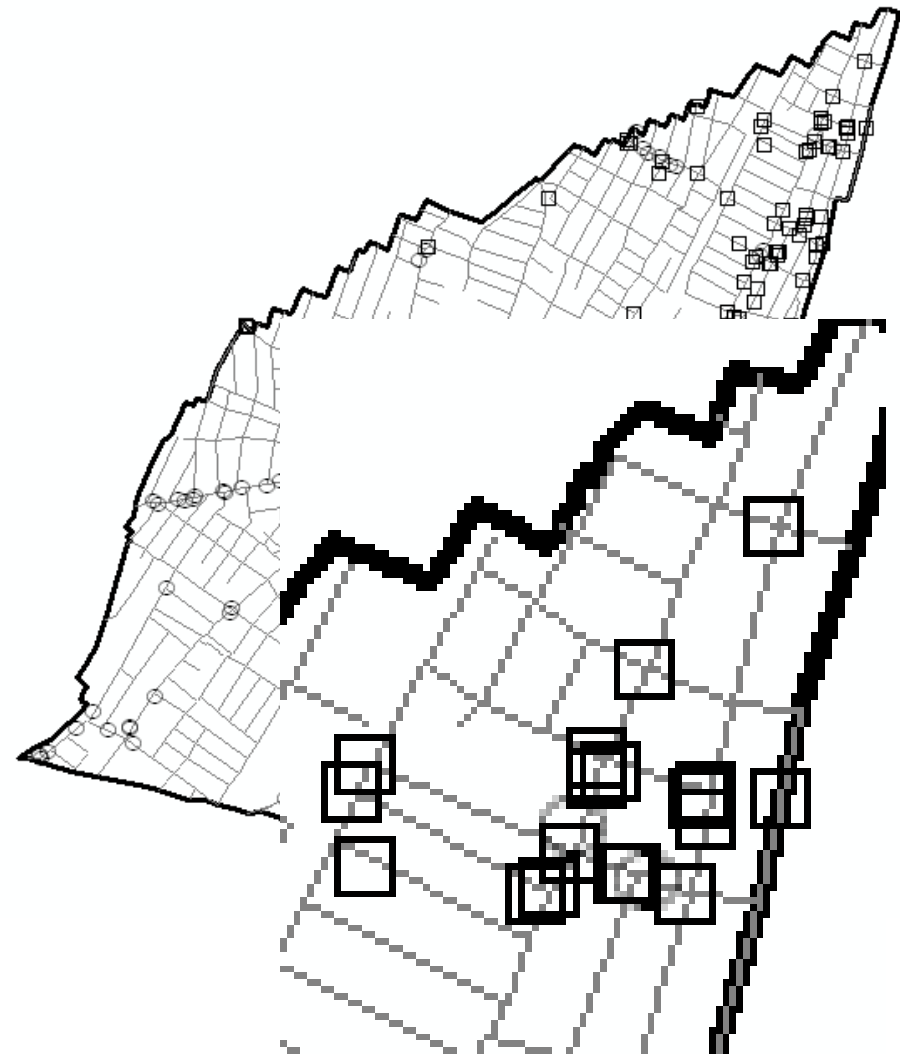
- ***STEP 6: Select particular risk factors to include in the risk terrain model.***
- All three risk factors identified in Step 5 will be included.
- Raw data in tabular form (i.e. Excel spreadsheets) was provided by the Township police and the many **datasets they maintain, validate and update regularly to support internal crime analysis and police investigations.**
- Attributes + **addresses** + time stamps + ??
- **State of the art of the investigation including the punishment and legal procedure.**



• **STEP 7: Operationalize risk factors to risk map layers.**

- The tabular data was geocoded to street centerlines of Irvington to create point features representing:
  - the locations of gang members' **residences** (hidden on the map to protect the gang members),
  - retail **business outlets**,
  - and **drug arrests**, respectively as three separate map layers.

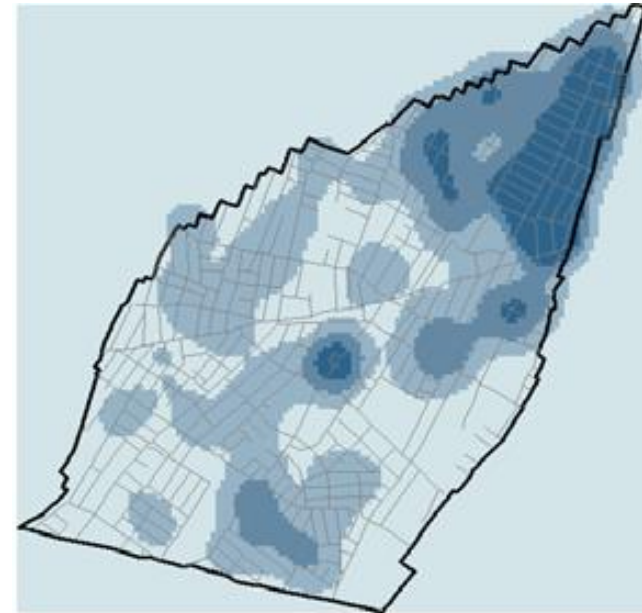
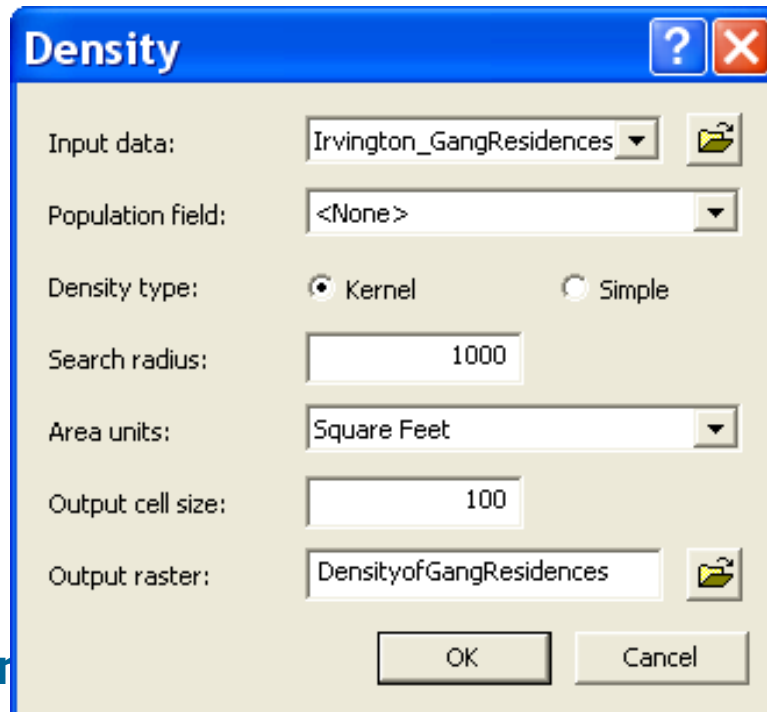
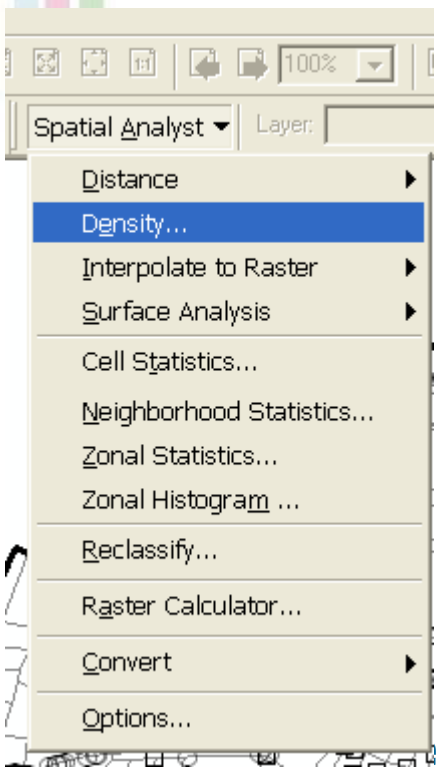
## Step 7





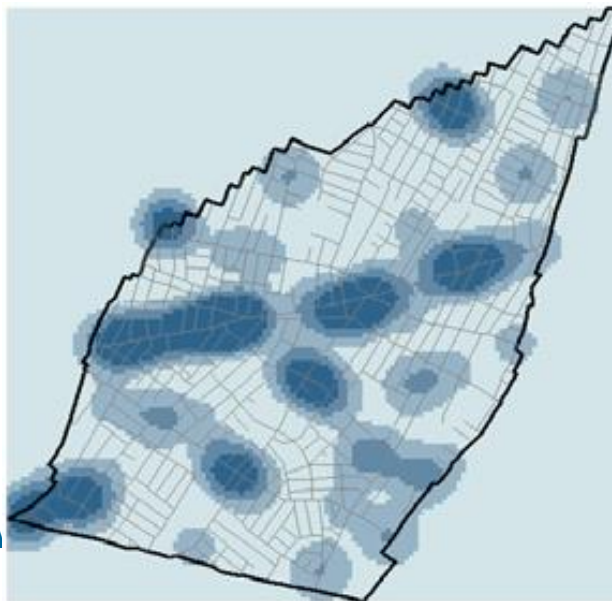
# Step 7a – gang member residence

**The spatial influence** of the “gang members’ residences” risk factor was operationalized as: “Areas with **greater concentrations of gang members residing will increase the risk of those places having shootings.**” So, a **density map** was created from the points of gang members’ residences. **Jádrové vyhlazování – proměnné ?**



## Step 7b - infrastructure

- The spatial influence of the “infrastructure” risk factor was operationalized as:
- “**High concentrations** of bars, strip clubs, bus stops, check cashing outlets, pawn shops, fast food restaurants, and liquor stores **will increase the risk** of those dense places having shootings.”



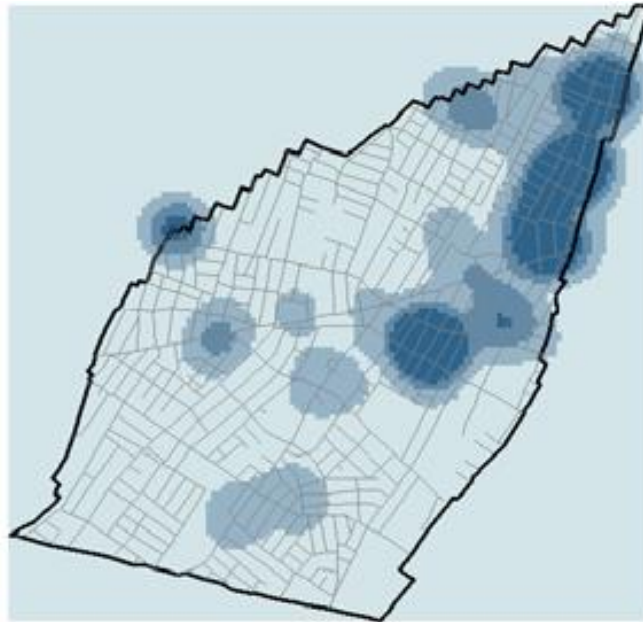




## Step 7C – the drug arrest

the “drug arrest” risk factor was operationalized as:

- “Areas with **high concentrations** of drug arrests **will be at a greater risk for shootings** because these arrests create new ‘open turf’ that other drug dealers fight over to control.”



Kartografické modelov



## Step 7 – map density method details

- **Kernel density** values were calculated for each of the risk map layers so that *points lying near the center of a cell's search area would be weighted more heavily than those lying near the edge*, in effect smoothing the distribution of values.
- Cells within each density map layer were **classified into four groups according to standard deviational breaks**. The dark blue colored cells had values in the **top five percent** of the distribution and were considered the “**highest risk**” places.

# Step 7d – distance from infrastructure

- The spatial influence of the “infrastructure” risk factor was also operationalized as:
- “The **distance of one block**, or about 350ft (app. 100 m), from a facility poses the greatest risk of shootings because **victims** are often **targeted** when **arriving** at or **leaving** the establishment.”

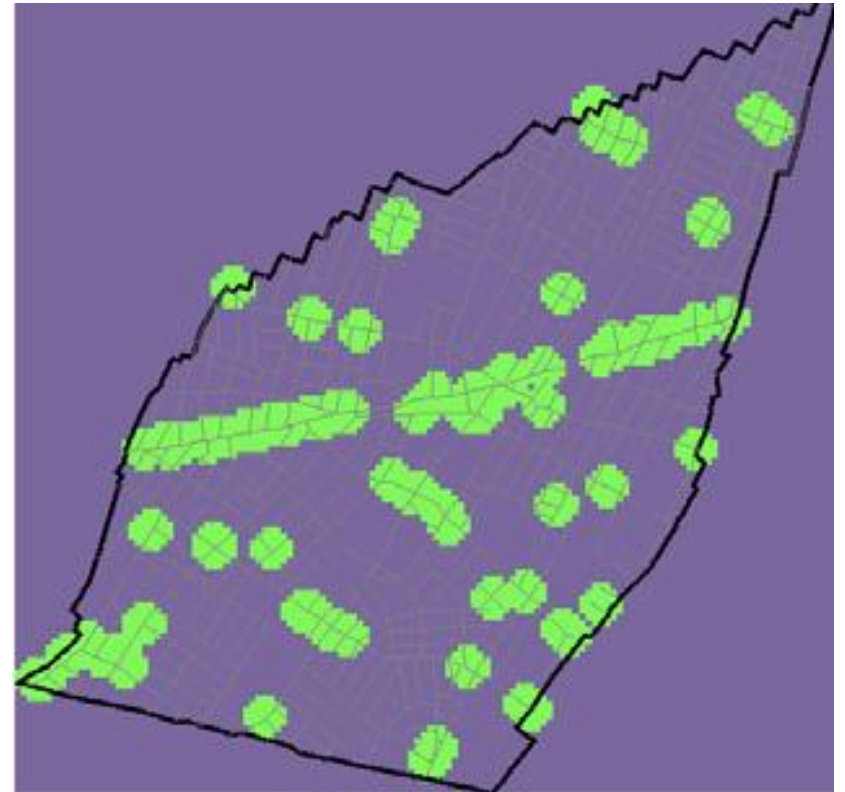
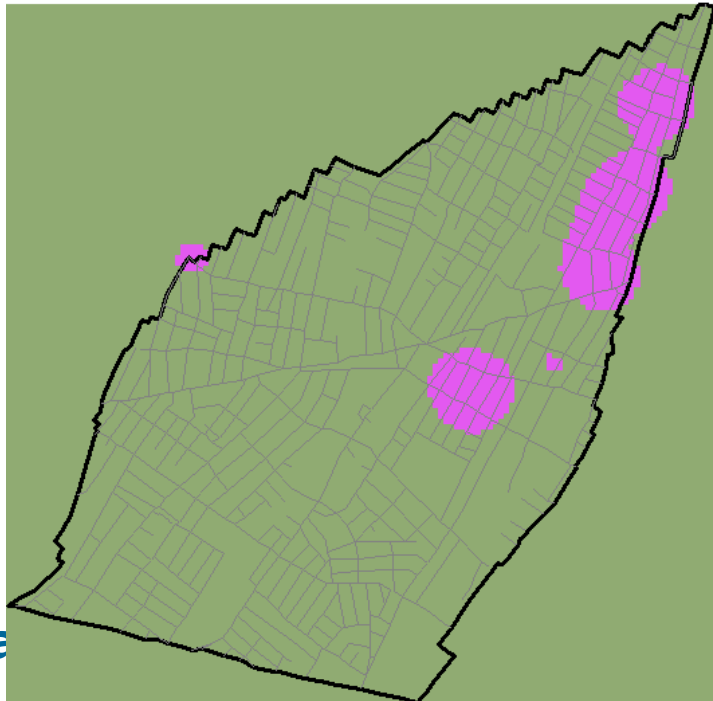
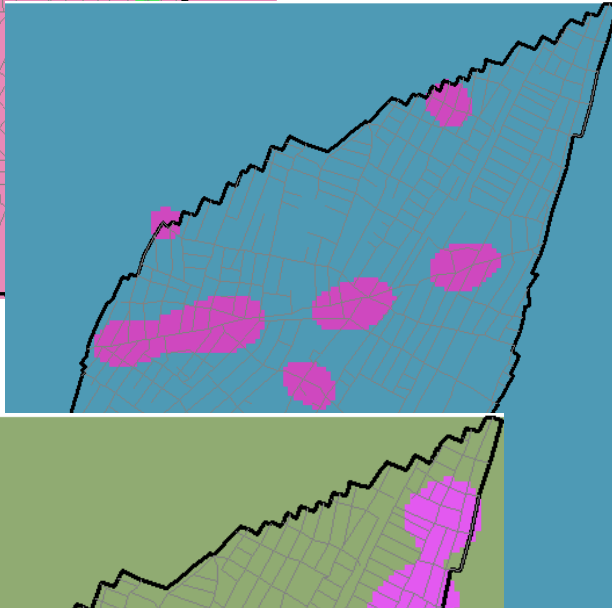
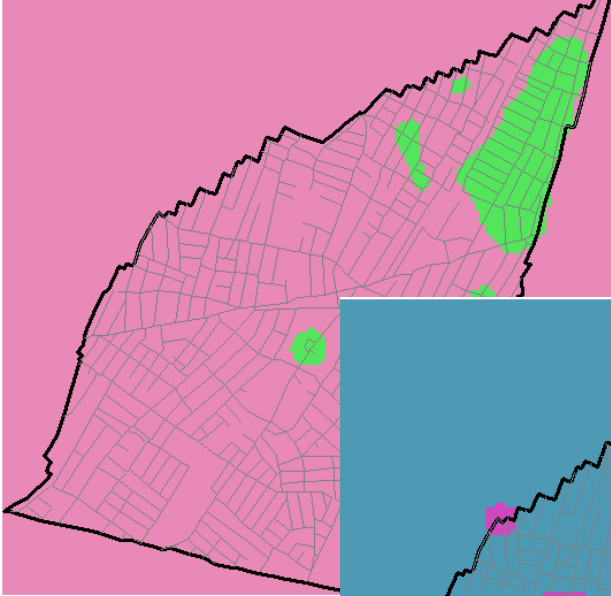


The slide features a logo in the top-left corner consisting of a stylized globe with blue and white lines, and the letters 'IGC' below it. The background is decorated with numerous vertical bars of various colors (blue, green, yellow, orange, pink, purple) of varying heights, creating a bar chart-like effect.

## 7e – final operationalization

- **We** are only interested in knowing where places are the most at risk for shootings, so we used a **binary-valued schema** to designate the “**highest risk**” places across all four risk map layers.
- The highest risk places of each risk map layer, respectively, will be given a value of “1”; all other places will be given a value of “0”.
- All risk factors are operationalized as **aggravating factors**, so these values will **remain positive**.

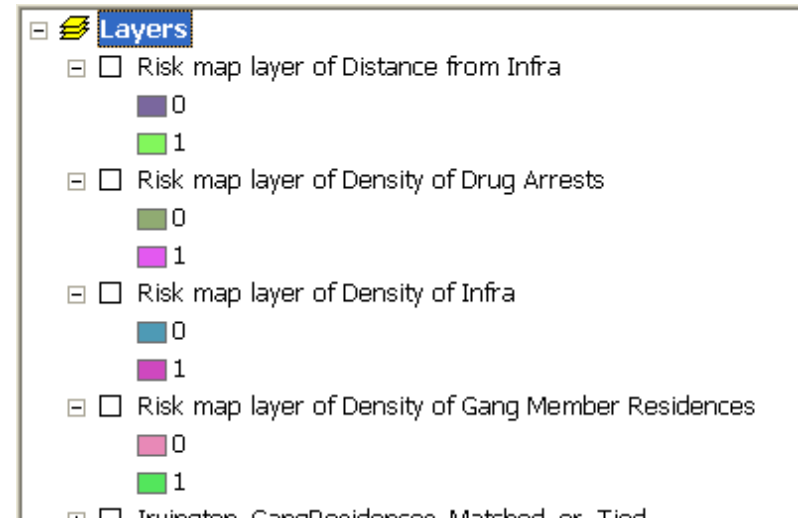
# Step 7 - reclassification





# Step 7 – final comparison

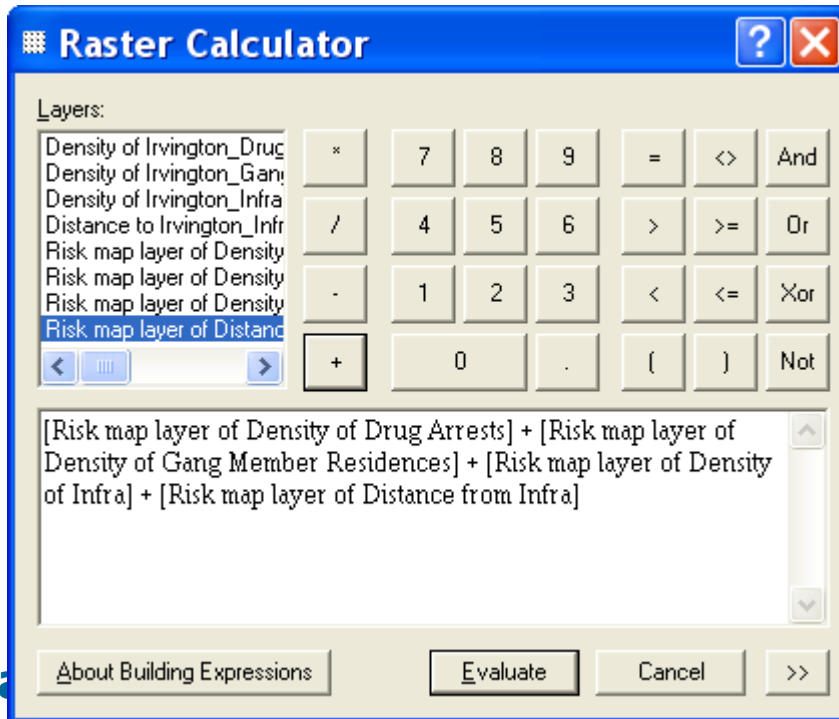
- We now have **four (final) risk map layers, operationalized from three risk factors.**
- **Binary** reclassification – 0 – 1
- The cells of different map layers are the same size and were classified in a standard way, the risk **map layers can be summed together** to form a **composite risk terrain map.**





# Step 8 + 9 - Inter Risk Map Layer Weighting and CRTM

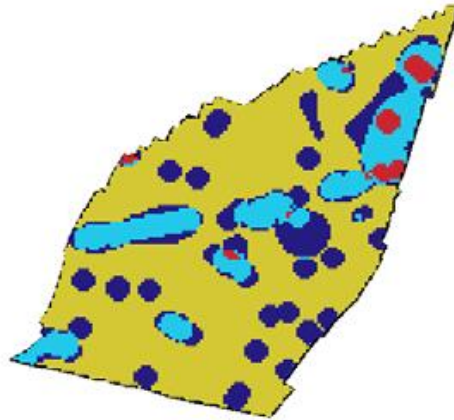
All risk **map layers** will carry equal weights to produce an **un-weighted risk terrain model**. It is assumed, for example, that being in a place with a high concentration of drug arrests **poses the same risk** of having a shooting as being in a place with a high concentration of gang member residences. Unless we know better 😊 !!



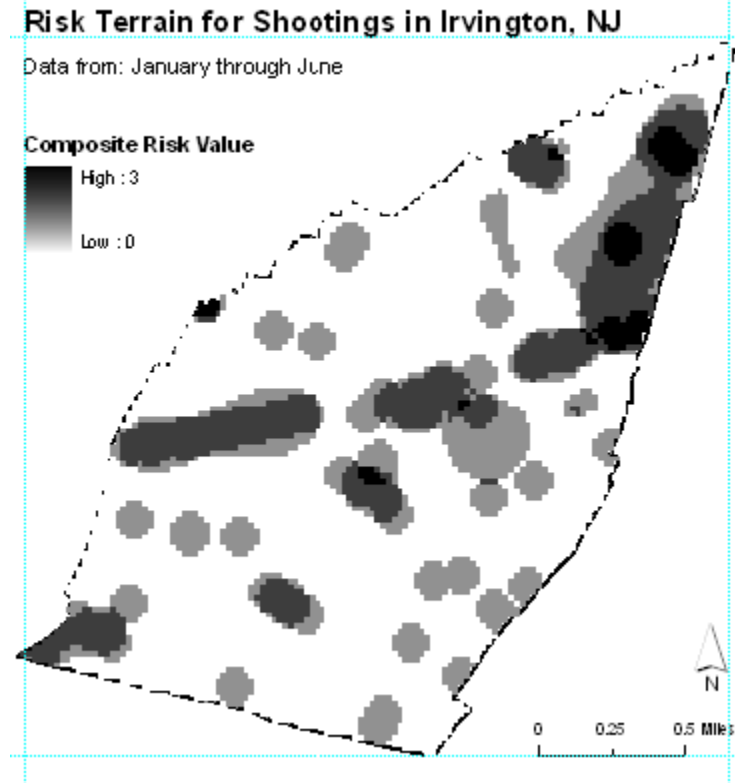


# STEP 10 - Finalize the Risk Terrain Map to Communicate Meaningful Information.

- Clip our risk terrain map to the boundary of Irvington.



- produce a final map with shades of grey and layout.





## Step 10 – make the risk count

- convert the risk terrain map from raster to vector we can (still using the regular structure converted to square polygons):
- **count the number of shootings that actually occur in the high-risk areas during the subsequent time period;**
- calculate the **square area** of the highest risk areas (i.e., places with a composite risk value of 3);



# Step 10 – make the risk count

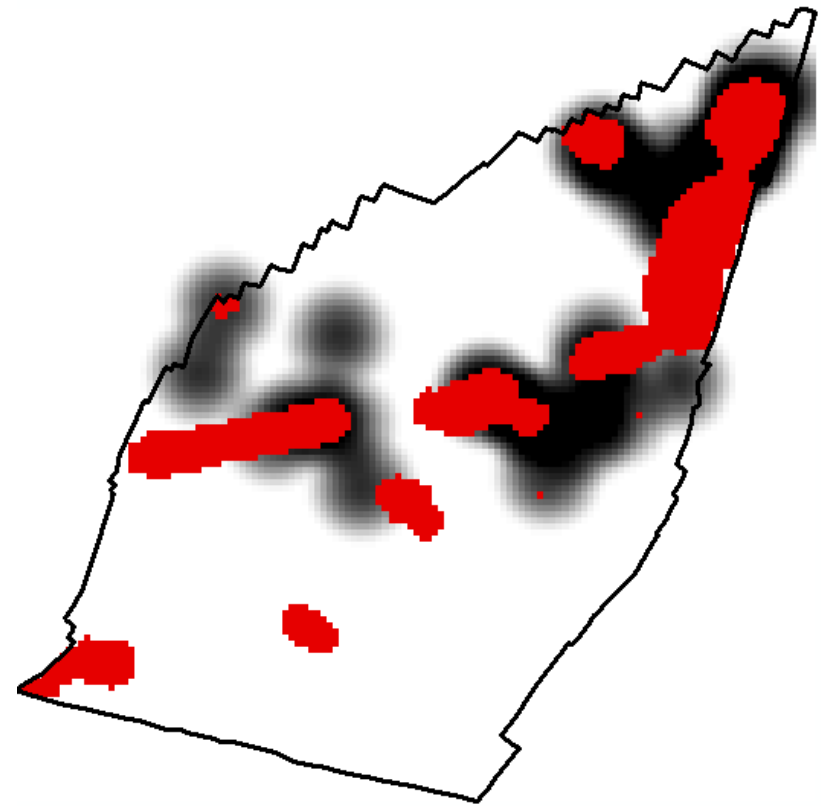
- Select all street segments within these areas to inform police commanders about where patrols might be increased.
- Operationalise the command and control on the day by day basis.

The screenshot shows the ArcMap interface with a map of street segments. The 'Layers' panel on the left lists various data layers, including 'Composite Risk Value' and 'Irvington\_Roads\_Clippped'. The 'Selected Attributes of Irvington\_Roads\_Clippped' table is displayed at the bottom, showing a list of street segments with their attributes.

FID	Shape*	TLID	RNODE	INODE	LENGTH	FENAME	FETYPE	FEDIRS	CFCC	FRADDL	TOADDL	FRADDR	TOADDR	ZIPL	ZIPR	CFCC1	CFCC2	SOURCE	COUNTRY
432	Polyline	63464488	8695	8434	0.14857	18th Ave	A41	376	438	395	399	07111	07111	A	A4	A	A	ESSEX	
436	Polyline	63464492	8775	8695	0.04408	18th Ave	A41	354	374	335	353	07111	07111	A	A4	A	A	ESSEX	
877	Polyline	63465432	8842	8775	0.03914	18th Ave	A41	344	352	343	353	07111	07111	A	A4	A	A	ESSEX	
878	Polyline	63465433	8925	8842	0.0476	18th Ave	A41	328	342	327	341	07111	07111	A	A4	A	A	ESSEX	
447	Polyline	63464503	8396	8127	0.14332	19th Ave	A41	171	235	172	234	07111	07111	A	A4	A	A	ESSEX	
470	Polyline	63464527	8460	8396	0.04208	19th Ave	A41	161	169	156	170	07111	07111	A	A4	A	A	ESSEX	
472	Polyline	63464529	8547	8460	0.0546	19th Ave	A41	141	159	140	154	07111	07111	A	A4	A	A	ESSEX	
489	Polyline	63464553	8573	8522	0.10016	21st St	A41	372	410	371	411	07111	07111	A	A4	A	A	ESSEX	
840	Polyline	63465009	8477	8490	0.02137	21st St	A41	0	0	413	417	07111	07111	A	A4	A	A	ESSEX	

# RTM validation

- **Comparison with the subsequent time period (June 1 – December 31) – high risk RTM classes and hot spot analysis of actual shooting accidents.**
- About 50% (15 out of 31) of the shootings during the subsequent time period (July 1 to December 31) happened in these high-risk cluster areas.



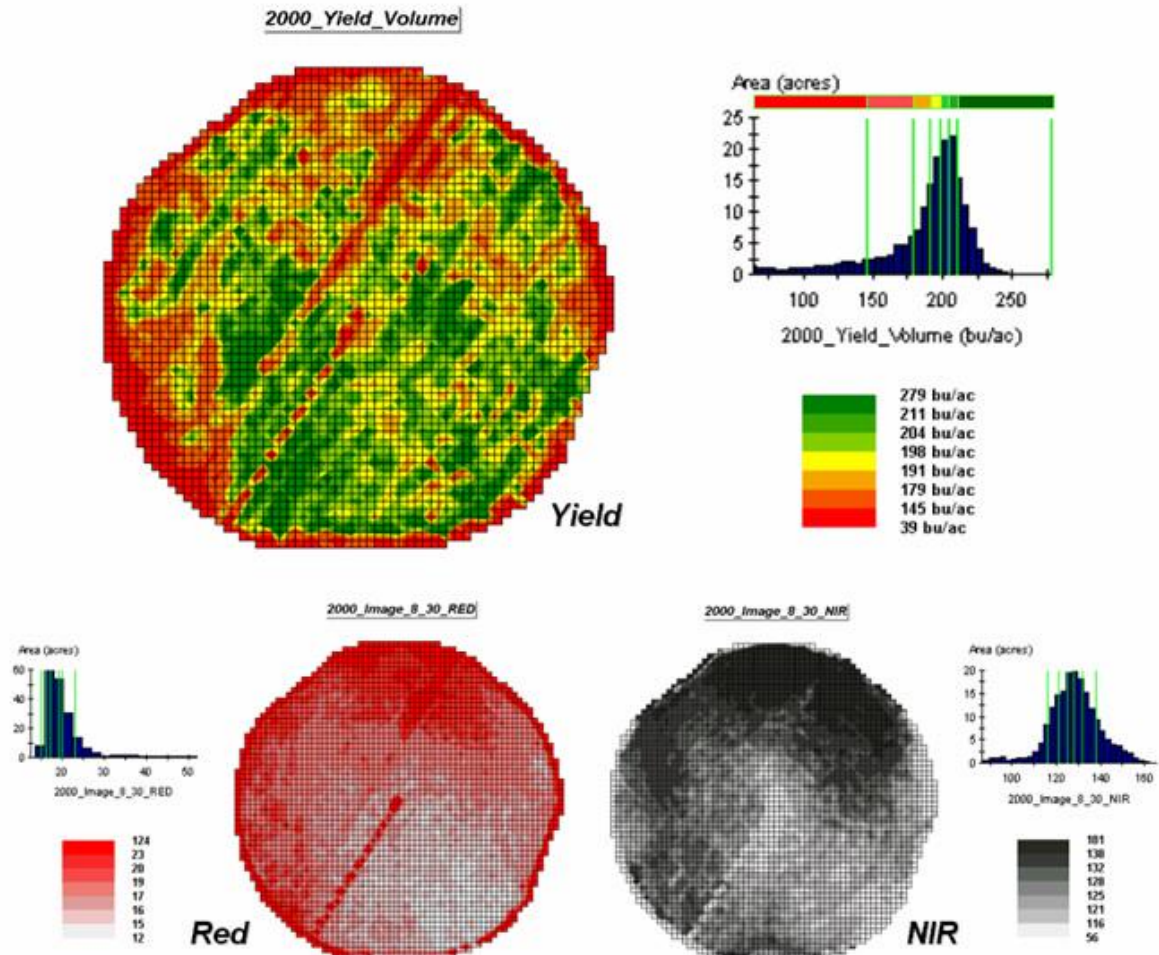
# Things to remember

- **Remember**, risk terrain modeling is only a *tool for spatial risk assessment*; it is not the solution to crime problems.
- You (the analyst) give **value and meaning to RTM**, so be innovative in your thinking about risk factors and how risk terrain maps can be applied to police operations.

# Případová studie

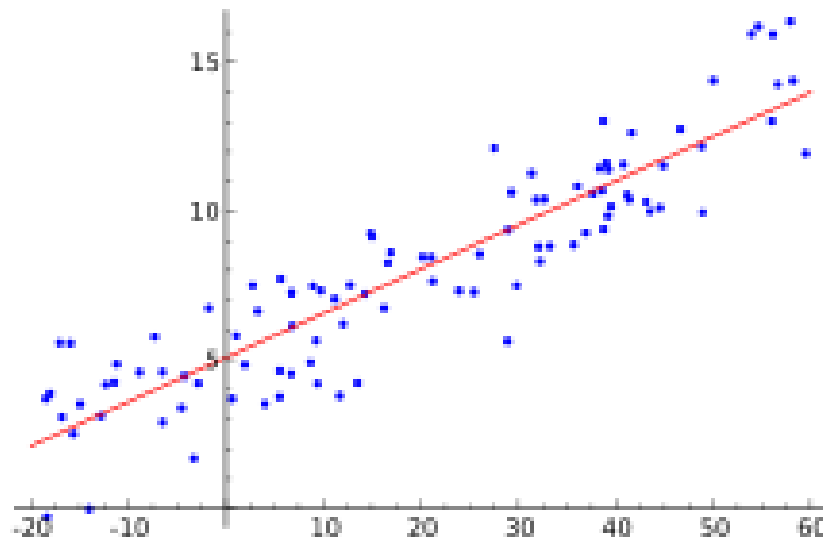
- **Využití prediktivního modelování pro precizní zemědělství (J. Berry).**
- **Výnosy kukuřice** – nízké (39 – červená) . Vysoké (279 – zelená) – **závislá proměnná** identifikující ten fenoméne (jev), který chceme predikovat.
- **Nezávislé proměnné** jsou použity pro to, aby bylo možné odhalit prostorové vztahy a vytvořit predikční rovnici.
- Využití data DPZ – **odrazivost povrchu rostlin** v červené a části spektra (RED) a v části blížíící se infračervené (NIR).

## Kartografické modelování



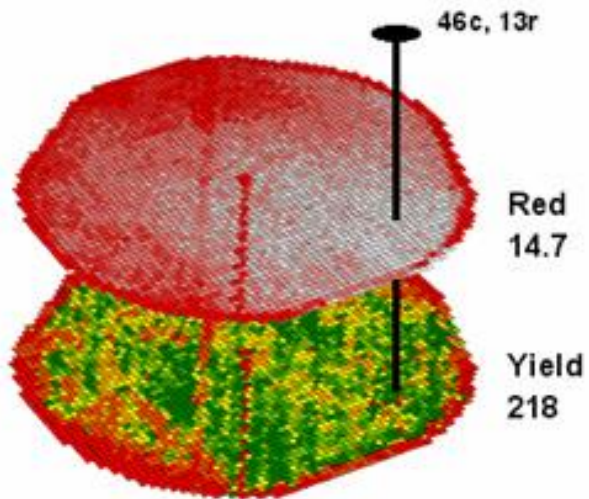
# Případová studie II

- Korelační diagram (Scatter plot) pro všechny existující dvojice hodnot.
- Predikční rovnice vytvořená pomocí regresní analýzy – křivka nejlépe charakterizující datové rozložení.

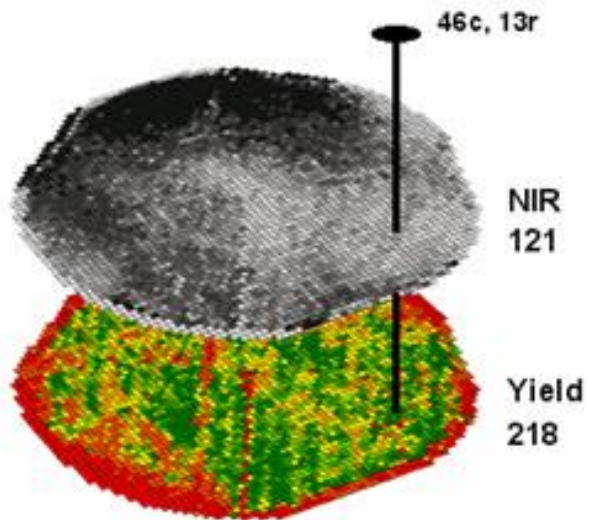
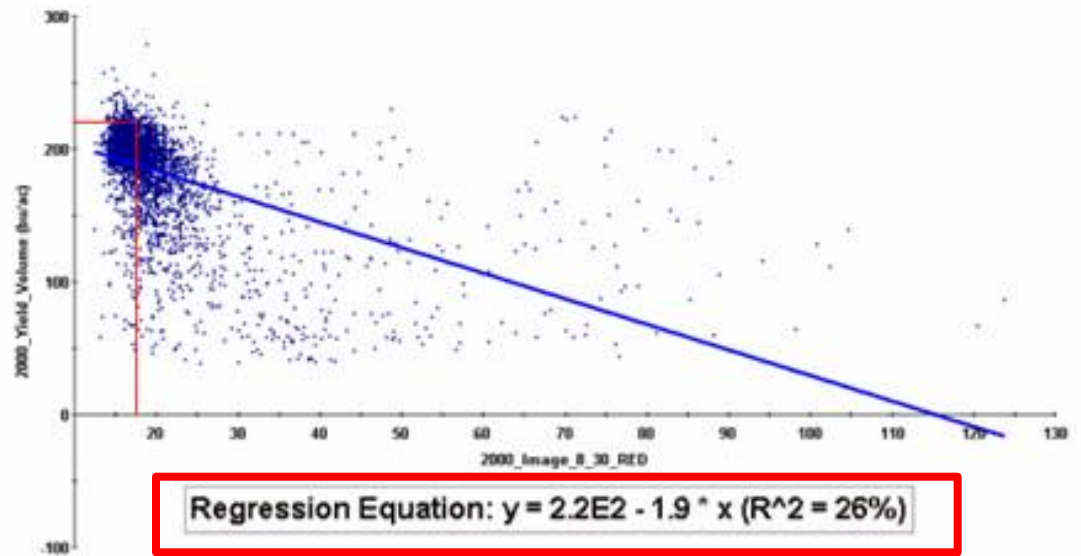


- Využití predikční rovnice pro další lokality.  
Kartografické modelování

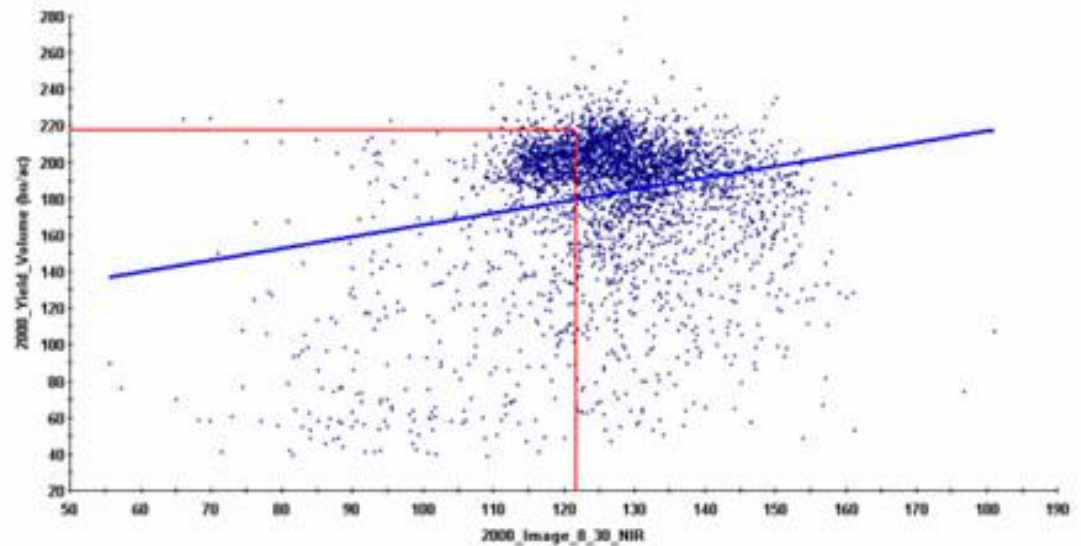




Graph of 2000\_Yield\_Volume vs 2000\_Image\_8\_30\_RED



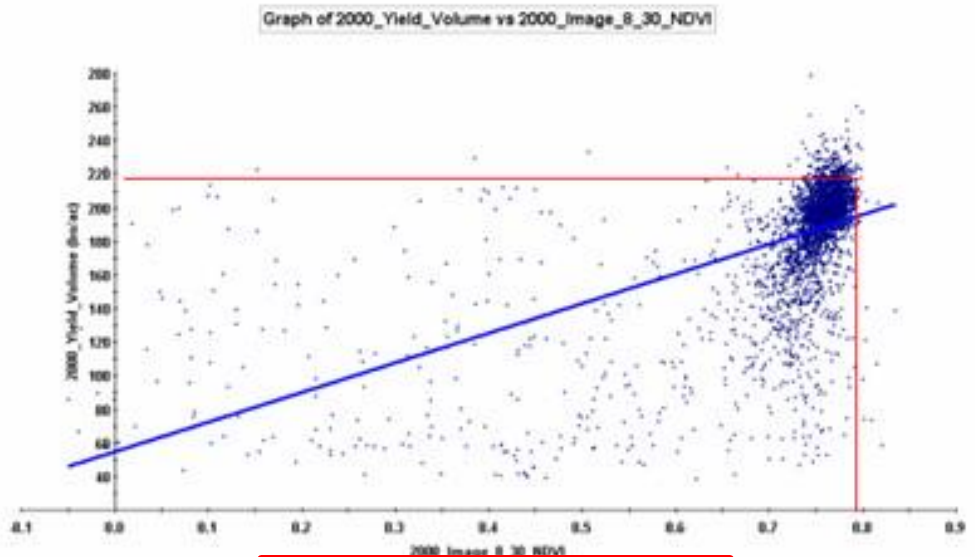
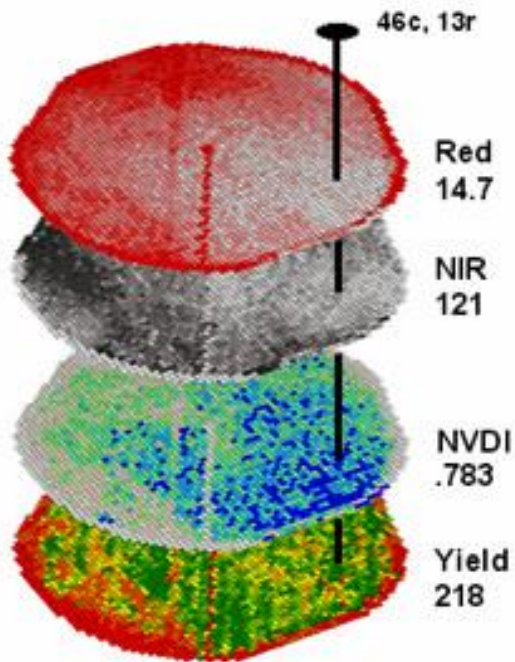
Graph of 2000\_Yield\_Volume vs 2000\_Image\_8\_30\_NIR



# Případová studie IV

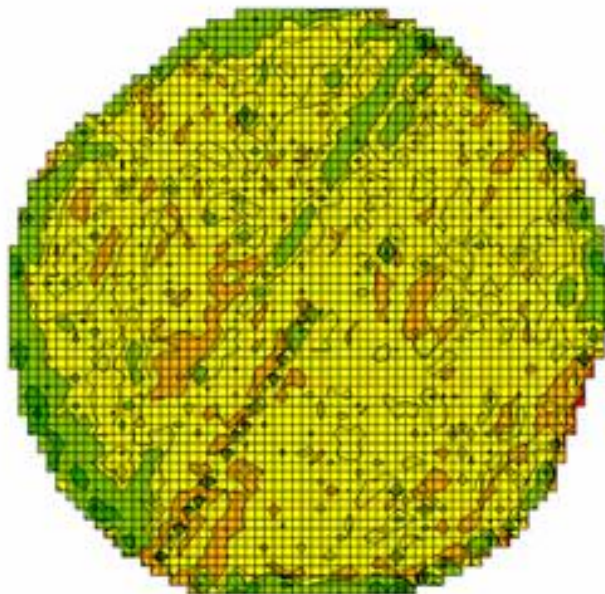
- **Problém?**
- **Predikční křivka nevystihuje rozdělení dat - nízké hodnoty  $R^2$  (jaké hodnoty jsou vyhovující?)**
- **Možnosti využití kombinovaného indexu - NDVI**
- **Normalized Density Vegetation Index (NDVI)**
- **$NDVI = ((NIR - Red) / (NIR + Red))$**
- **Srovnání predikované a skutečné hodnoty (kalibrace modelu) - mapa odchylek.**
- **Průměrná chyba 2,26 q/ha.**



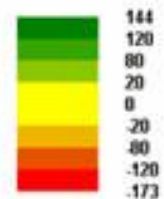


Regression Equation:  $y = 55 + 1.8E2 \cdot x$  ( $R^2 = 30\%$ )

Error\_Map

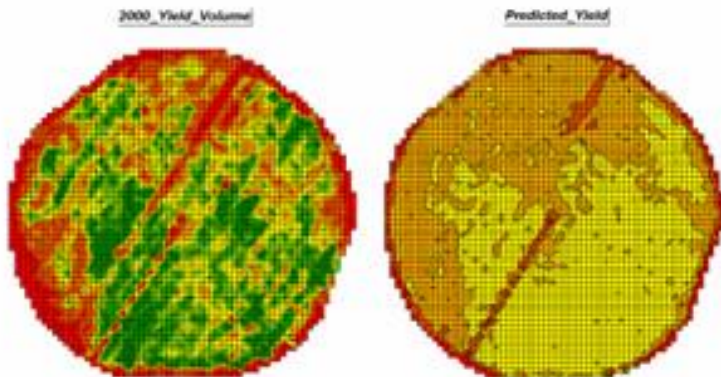


Statistics	
Min:	-173
Max:	144
Range:	318
Mean:	2.62
Median:	-4.23
Std. Dev.:	32
Variance:	1,025
Gridded Area:	189 acres



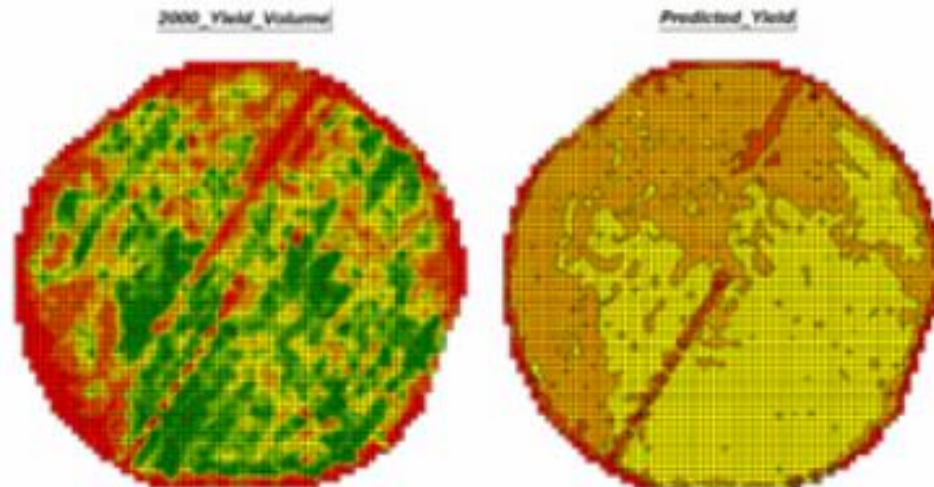
67%

Regression Equation:  $y = 55 + 1.8E2 \cdot x$  ( $R^2 = 30\%$ )



# Případová studie IV

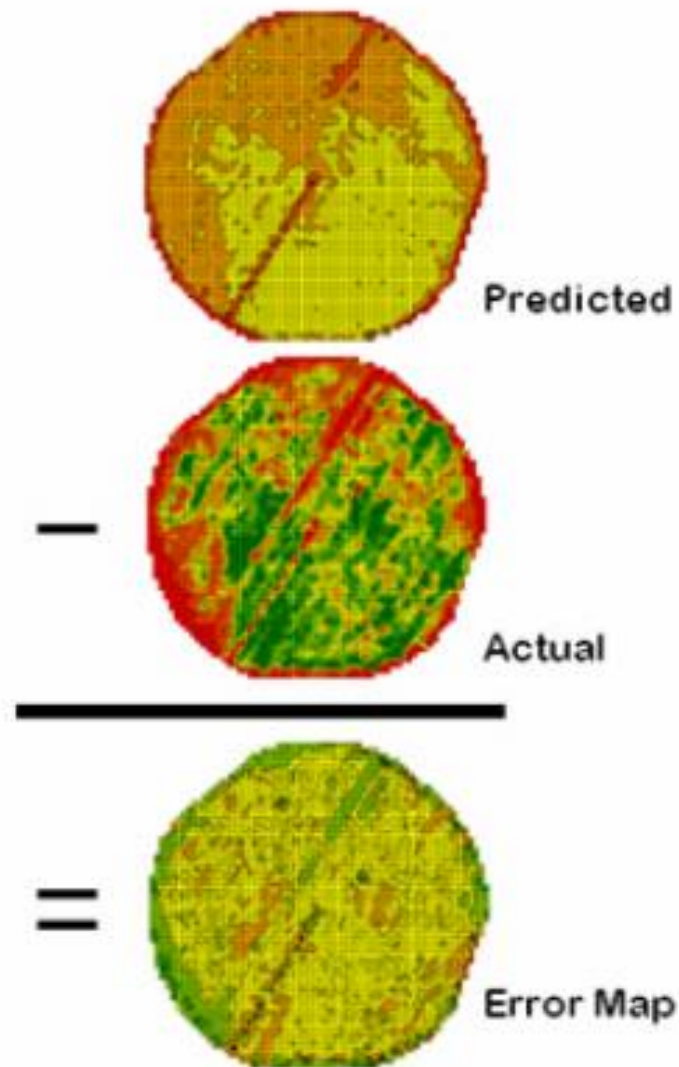
- **Výsledky shlazují skutečné výnosy.**



- **Nejedná se o skutečnou kalibraci modelu, ale spíše o první zjednodušený náhled, jaké by mohly výnosy být.**
- **Jak můžeme výsledný model dále zlepšit?**

# Případová studie V

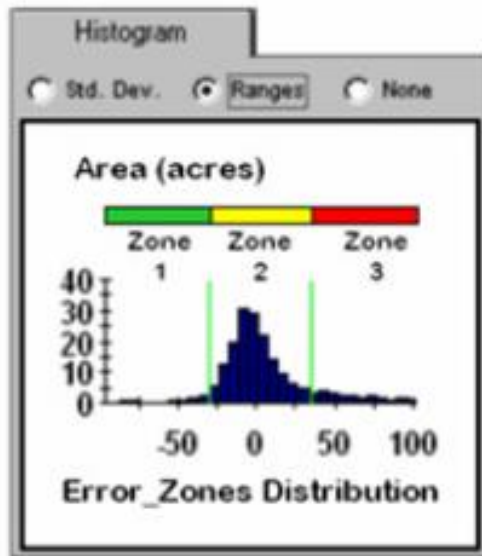
- **Bližší pohled na mapu odchylek**
- **Průměrná chyba 2,62 q/ha.**
- **67% odhadu  $\pm$  20 q/ha.**
- **ALE – některé lokality až +144 a -173 q/ha.**



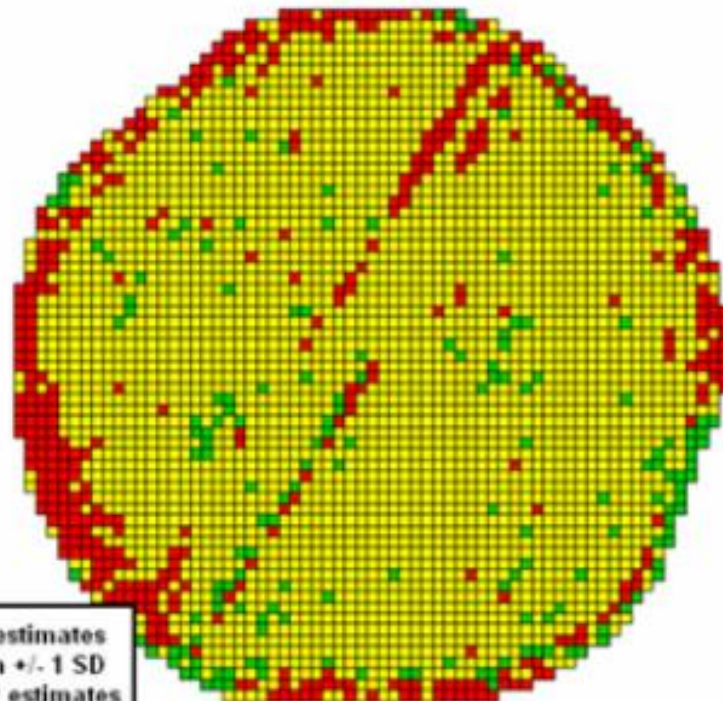


# Případová studie VI

- Řešení?
- Stratifikace datové sady – rozdělení do skupin ze stejnými charakteristikami.



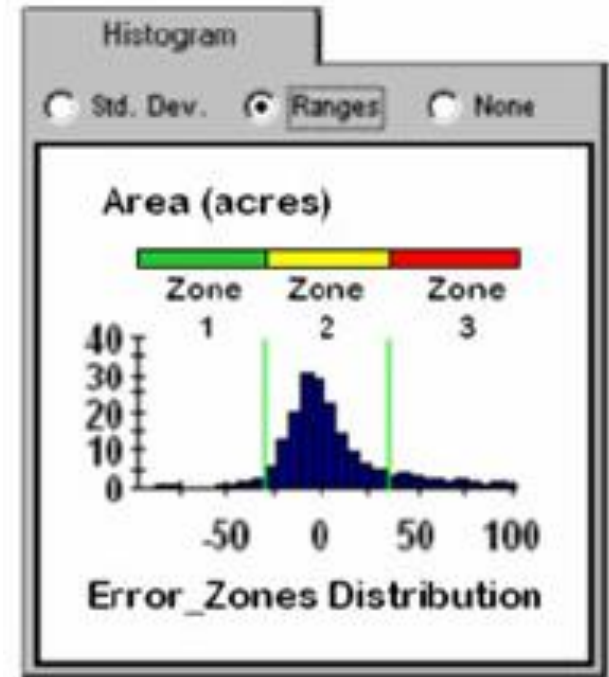
Error\_zones



# Případová studie VII

- Predikční rovnice bude lépe vystihovat jednotlivé vrstvy, než jedna rovnice pro celá data.
- Více technik pro stratifikaci.
- Využití histogramu – plus/minus směrodatná odchylka dělí histogram na 3 zóny.
- Predikce funguje pro zónu 2.
- Pro zóny 1 a 3 jsou výsledky pod a nadhodnocené.
- Specifická predikční rovnice pro každou zónu by měla dát lepší predikci.

Kartografické modelování

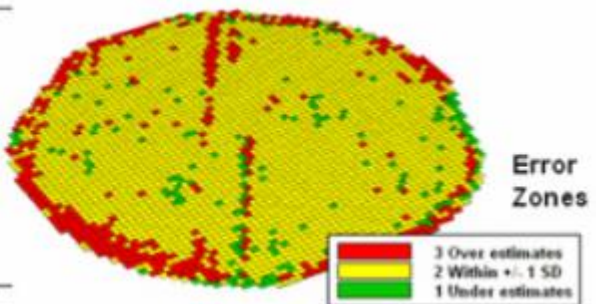




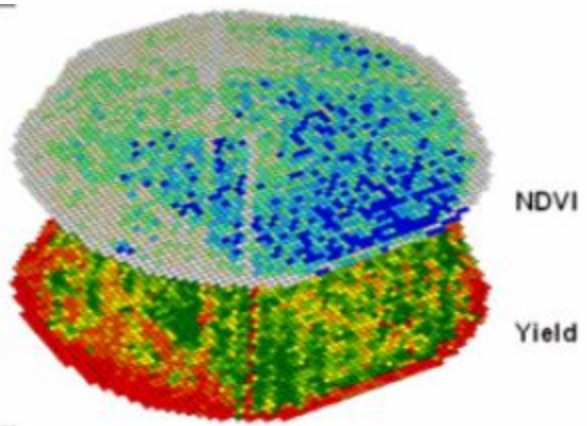
# Výpočet predikce pomocí separovaných predikčních rovnic

- **Vstupy – NDVI a výnosová mapa.**
- **Algoritmus nejdříve zkontroluje mapu odchylek a určí, do které ze 3 zón daná oblast patří.**
- **Následně jsou použity regresní rovnice pro predikci po zónách.**
- **Složená predikční mapa vytvořena pomocí 3 rovnic a NDVI dat.**

Template Map

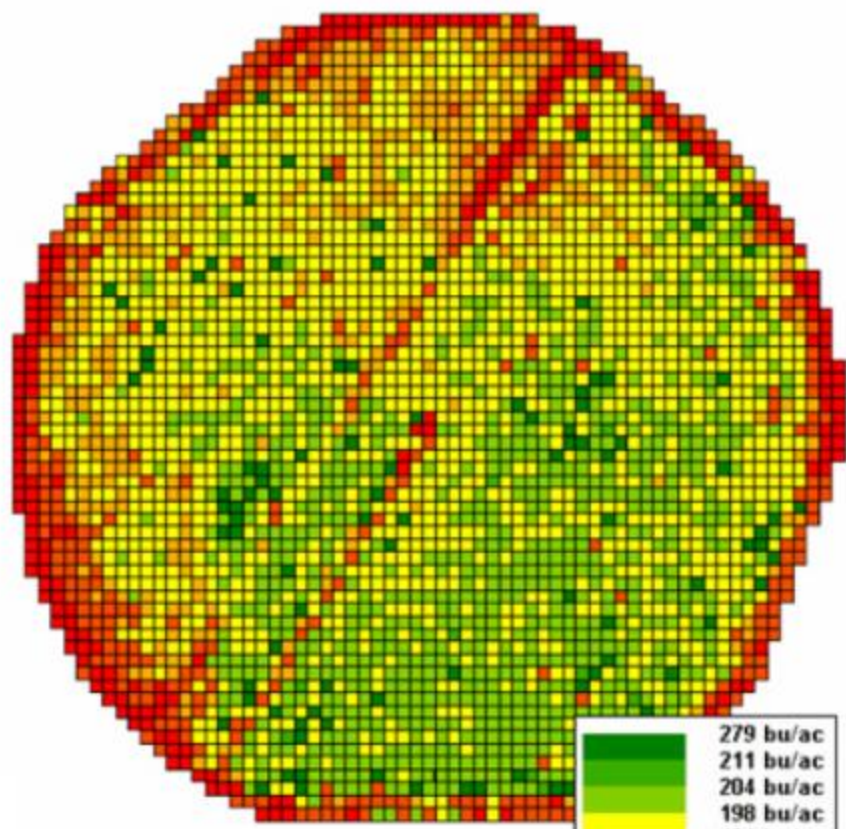


Data Maps

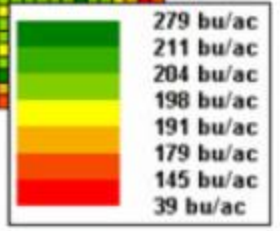


Prediction Equations

- Zone 1 ■  $Y = 145.40 + 110.79X$  ( $R^2 = .68$ )
- Zone 2 ■  $Y = 32.93 + 215.06X$  ( $R^2 = .60$ )
- Zone 3 ■  $Y = -4.85 + 169.38X$  ( $R^2 = .42$ )



*Prediction\_composite*



# Kartografické modelování



# Případová studie VIII

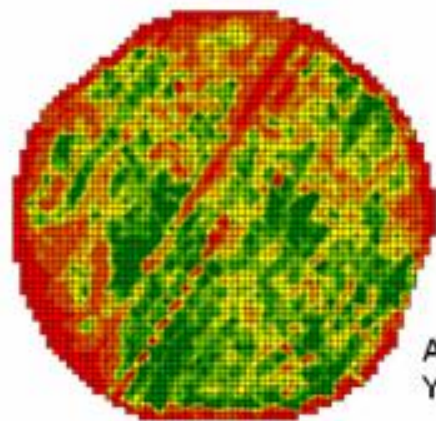
- **Vizuální srovnání :**
  - Skutečné výnosy
  - Predikovaná mapa pro celou oblast
  - Predikovaná stratifikovaná mapa
- **Mapa odchylek pro stratifikovanou predikci – 80% odhadu je  $\pm 20$  q/ha.**
- **Průměrná chyba je pouze 4q/ha.**
- **Dobrá predikce úrody na základě DPZ více jak měsíc před sklizní 😊**

Min [ >= ]	Max [ < ]	Count	acres	% Gridded Area	Color
120	144	0	0	0	Dark Green
80	120	30	1.72	0.91	Light Green
20	80	477	27.4	15	Yellow-Green
0	20	1236	70.9	38	Yellow
-20	0	1380	79.2	42	Orange
-80	-20	165	9.47	5	Red-Orange
-120	-80	1	0.0574	0.03	Red
-173	-120	0	0	0	Dark Red

80%

Statistics	
Min:	-81.2
Max:	113
Range:	194
Mean:	3.97
Median:	1.12
Std. Dev.:	19.2
Variance:	369
Gridded Area:	189 acres

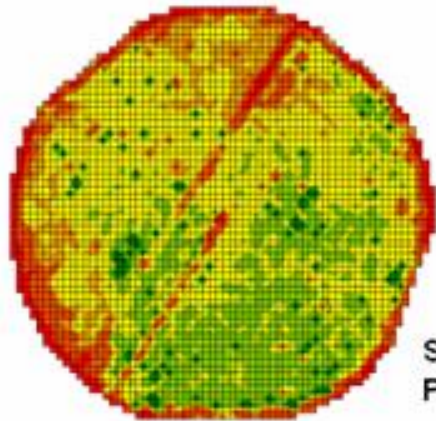




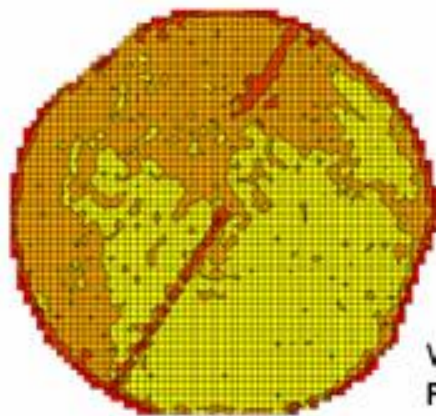
Actual Yield



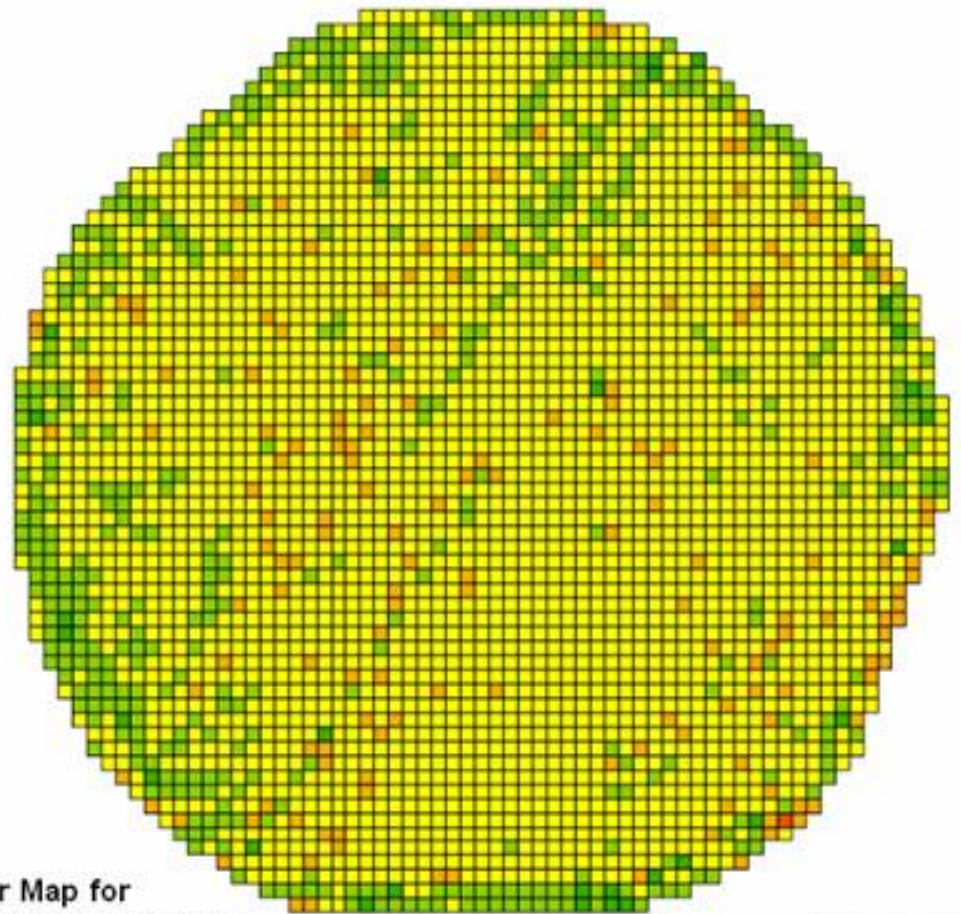
279 bu/ac  
243 bu/ac  
209 bu/ac  
175 bu/ac  
141 bu/ac  
107 bu/ac  
73 bu/ac  
39 bu/ac



Stratified Prediction



Whole Field Prediction



Error Map for Stratified Prediction

Min [ >= ]	Max [ < ]	Count	acres	% Gridded Area	Color
120	144	0	0	0	Dark Green
80	120	30	1.72	0.91	Green
20	80	477	27.4	15	Light Green
0	20	1236	70.9	38	Yellow-Green
-20	0	1380	79.2	42	Yellow
-80	-20	165	9.47	5	Orange
-120	-80	1	0.0574	0.03	Red-Orange
-173	-120	0	0	0	Red

} 80%

Statistics

Min: -81.2  
Max: 113  
Range: 194  
Mean: 3.97  
Median: 1.12  
Std. Dev: 19.2  
Variance: 369  
Gridded Area: 189 acres

# Na co dávat pozor?

- **Odlišné způsoby stratifikace dat:**
  - *Prostorové zóny* – blízkost či vzdálenost od okraje pozemku.
  - *Závislé mapové zóny* – oblasti různé intenzity výnosů.
  - *Datové zóny* – půdní druhy, zrnitost, nutriční hodnoty.
  - *Korelační mapové zóny* – mikroreliéf – hřbety a deprese.

# Na co dávat pozor?

- Citlivé užití mapy odchylek – zejména pro další **extrapolaci v čase a prostoru**.
- Nutno využít dalších oblastí pro **kalibraci a validaci**.
- V oblasti precizního zemědělství například možnost kombinace detailního měření v relativně dlouhé periodě (DPZ) s častým měřením s omezeným prostorovým výskytem (vzorky, senzory).



# Prediktivní modelování v ArcGIS

- **Obvykle se jednotlivé procedury modelování spouští samostatně a opakovaně.**

## **Možnost využít ModelBuilderu pro:**

- 1) Zaznamenání všech **postupných kroků** v modelování;
- 2) Snadná **opakovatelnost** modelování a **sdílení** s dalšími uživateli;
- 3) Lepší **vizuální reprezentace**, která vede k lepšímu pochopení celého průběhu modelování..



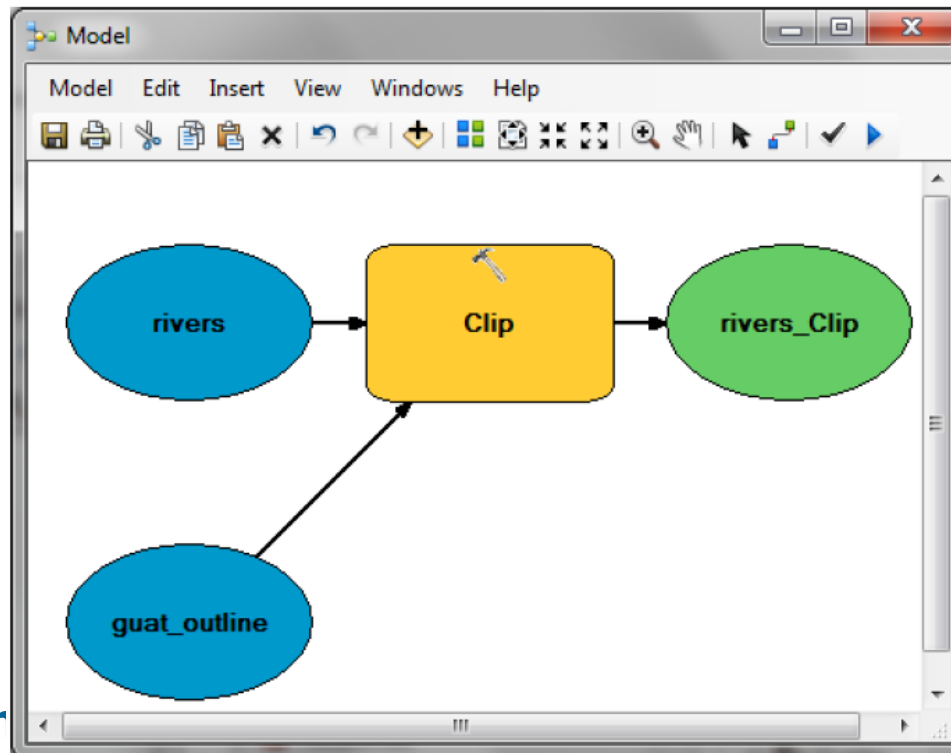
# Prediktivní modelování archeologického naleziště

- Prediktivní modelování v archeologii – „nástroj pro vyjádření pravděpodobnosti výskytu archeologického naleziště kdekoliv v krajině“.
- Snaha určit pravidla a preference pro výběr lokality danou kulturou.
- Zahrnuje deskriptivní analýzu přírodních faktorů pro známé lokality a snahu najít společné opakující se kombinace.
- Příklad: vybraná kultura (Mayové) preferovala historicky známá místa v **blízkosti** oceánu a mokřadů s výskytem porostů endemita *Salvia apiana*.
- Která místa ve zkoumané oblasti odpovídají podmínkám??



# 1. Omezení zkoumané oblasti

- Omezení oblasti na severní Guatemalu a oříznutí vybraných vodních toků pomocí funkce Clip.
- Vstupní a výstupní soubory + funkce.





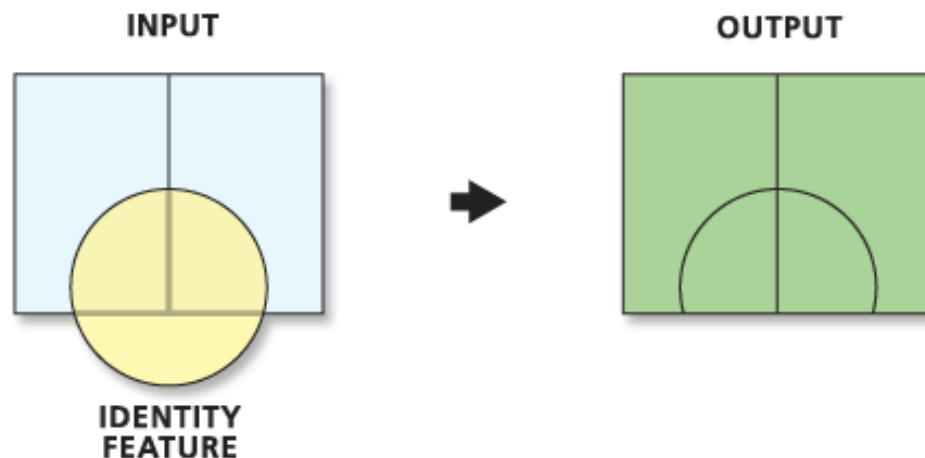


## 2. Změření vzdálenosti lokalit od řeky

- Říční síť nyní omezena na sledované území.
- Určení vzdálenosti potenciálních archeologických nalezišť od říční sítě – **Near**.
- Vyhledávací vzdálenost nastaveno na 5 km (=blízko).
- Všechny lokality blíže než 5 km mají určenou přesnou (vzdušnou) vzdálenost (**NEAR\_DIST**).
- Ostatní lokality mají přiřazenu hodnotu -1.

### 3. Kombinace přírodních podmínek

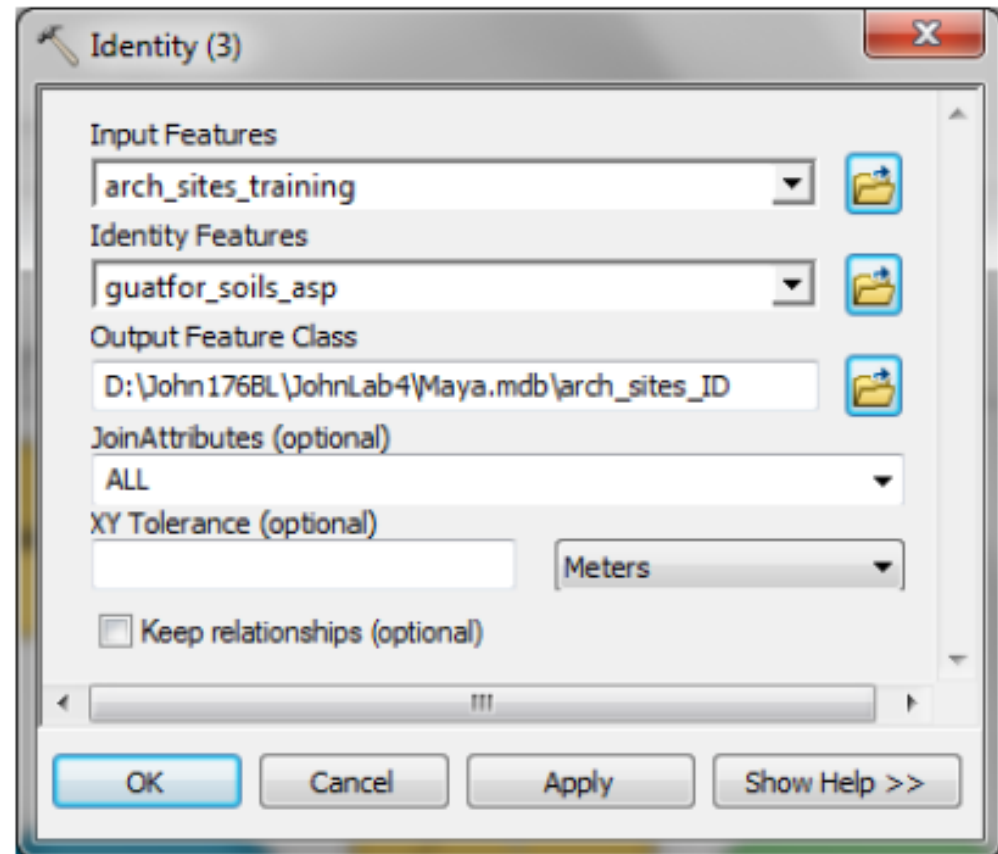
- Zjištění jaké přírodní podmínky obklopují naše archeologické lokality.
- **Vegetace – půdy – orientace svahu.**
- Nutná postupná analýza přírodních podmínek v několika krocích a postupné rozšíření atributové tabulky o přírodní ukazatele.
- Použití nástroje ***Identity***.
- Vegetace + půdy = PP1
- PP1 + orientace = PP2

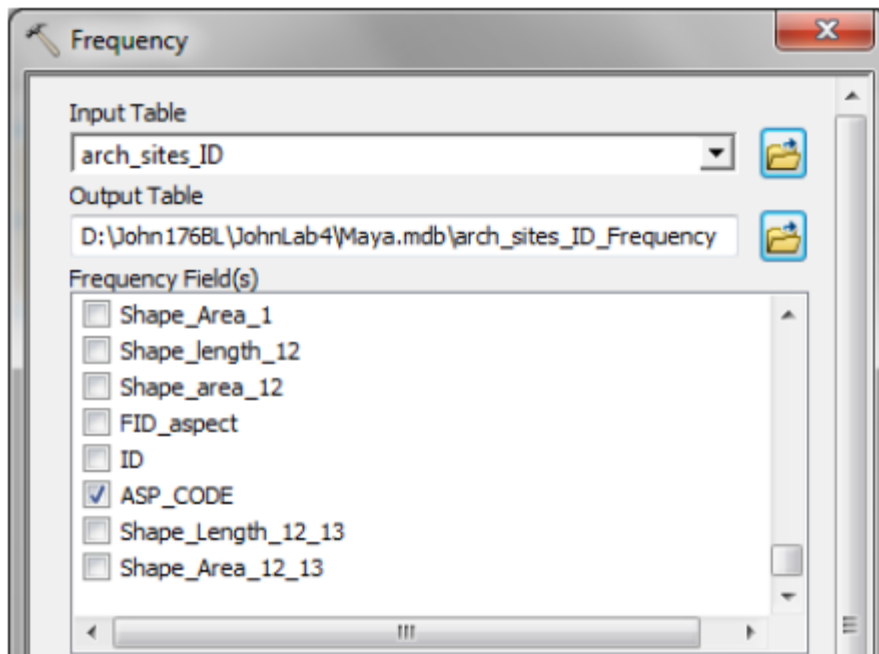




## 4. Přírodní podmínky pro archeologické lokality

- Spojení dat o archeologických lokalitách a PP2 pomocí nástroje **IDENTITY**.
- Následný výběr potřebných atributů z tabulky – nástroj Identity zachovává všechny atributy a vytváří další.
- Využití nástroje **Frequency**.

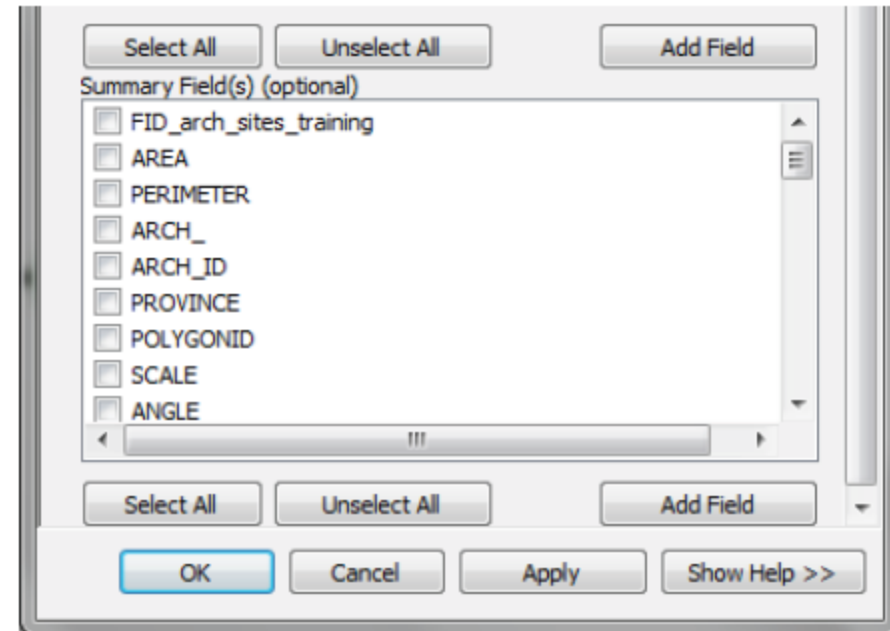




# Výběr atributů

- **Nutno zachovat:**
  - NEAR\_DIST - blízkost
  - DESC\_vegetace
  - R\_FERT - půda
  - ASP\_CODE - orientace

**Kartografické modelování**





# Finální model



Kartografické modelování

## 6. RUN a prozkoumání výsledků

- Určení hlavních shluků přírodních podmínek.
- Stanovení pracovních predikční hypotézy pro vybraná místa.
- Ověření hypotézy.

arch\_sites\_ID\_Frequency

OBJECTID *	FREQUENCY	NEAR_DIST	DESC_	R_FERT	ASP_CODE
1	1	-1	Inland swamp forest	4	10
2	1	-1	Lowland rain forest	1	2
3	2	-1	Lowland rain forest	1	5
4	2	-1	Lowland rain forest	1	9
5	1	-1	Non forest	1	10
6	1	68.570929		2	10
7	1	177.68938		2	9
8	1	274.989335	Lowland rain forest	1	4
9	1	327.802407	Non forest	1	8
10	1	427.268735	Inland swamp forest	2	4
11	1	546.290435	Non forest	1	7
12	1	593.566121	Lowland rain forest	4	6