APLIKACE ROBUSTNÍ VÁŽENÉ LOKÁLNÍ REGRESE NA TEPLOTNÍ ČASOVÉ ŘADY

MARIE BUDÍKOVÁ

1. Regresní model trendu

Zkoumáme teplotní časovou řadu Y_t, $t \geq 0$. Nechť máme k dispozici n pozorování této řady, přičemž i-té pozorování vyhovuje modelu

$$Y_i = f(t_i) + \varepsilon_i, \ i = 1, \ldots, n \ , \qquad (1)$$

kde $f(t)$ je neznámá trendová funkce a $\varepsilon_1, \ldots, \varepsilon_n$ jsou náhodné fluktuace pozorování $Y_1, \ldots, Y_n$ kolem trendové funkce. Předpokládáme, že $\varepsilon_1, \ldots, \varepsilon_n$ jsou nezávislé náhodné veličiny, $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{D}(\varepsilon_i) = \sigma^2 > 0$, $i = 1, \ldots, n$. Bez újmy na obecnosti budeme v teoretické části příspěvku předpokládat, že $t_i \in \langle 0, 1 \rangle$, $t_i = \frac{i-1}{n}$, $i = 1, \ldots, n$.

2. Neparametrické metody odhadu trendu

Neparametrický přístup k odhadu neznámé trendové funkce je založen na předpokladu, že $f(t)$ patří do třídy k-krát spojitě diferencovatelných funkcí C^k, $k \geq 0$, tedy funkce $f(t)$ může být v okolí bodu t_i vyjádřena polynomem

$$f(t_i) = \sum_{j=0}^{m} \alpha_j(t_i) t_i^j \ , \qquad i = 1, \ldots, n \ .$$

Odhad $f(t_i)$ polynomem provedeme nikoliv na celém intervalu $< 0, 1 >$, ale na lokálním intervalu $< t_i - b, t_i + b >$, kde $b \in (0, 1)$ je šířka regrese. Není-li interval $< t_i - b, t_i + b >$ subintervalem intervalu $< 0, 1 >$, odhad provedeme na subintervalu $< t_i - b, t_i + b > \cap < 0, 1 >$. K tomuto intervalu jsou příslušná pozorování $Y_{i-p}, Y_{i-p+1}, \ldots, Y_{i+q}$, kde $p = \min\{r, i-1\}$, $q = \min\{r, n-i\}$ a r je nejbližší celé číslo k součinu bn.

Sestavíme model lokální regrese

$$\begin{aligned} Y_{i+k} &= \sum_{j=0}^{m} \alpha_j(t_i) \left(\frac{i-1+k}{n} \right)^j + \varepsilon_{i+k} \\ &= \sum_{j=0}^{m} \beta_j(t_i) k^j + \varepsilon_{i+k}, \quad k = -p, -p+1, \ldots, q, \quad i = 1, \ldots, n \ . \end{aligned}$$

Metodou nejmenších čtverců dostaneme odhad $\hat{Y}_i = \hat{\beta}_0(t_i)$, který považujeme za odhad $\hat{f}(t_i)$ neznámé trendové funkce f v bodě t_i, $i = 1, \ldots, n$.

V případě, že při odhadu trendové funkce $f(t)$ nepředpokládáme, že všechna pozorování mají stejný vliv, můžeme zavést váhy w_i, $i = 1, \ldots, n$, které vystihují naši představu o vlivu jednotlivých pozorování Y_i, $i = 1, \ldots, n$ na odhad trendu a použít model vážené lokální regrese:

$$\begin{aligned} w_k Y_{i+k} &= \sum_{j=0}^{m} \beta_j(t_i) k^j \sqrt{w_k} + \varepsilon_{i+k} \ , \\ k &= -p, -p+1, \ldots, q, \quad i - 1, \ldots, n \ . \end{aligned}$$

Stejně jako v předešlém případě dostaneme $\hat{Y}_i = \hat{\beta}_0(t_i) = \hat{f}(t_i)$, $i = 1, \ldots, n$.

Další zlepšení neparametrických odhadů představuje robustní vážená lokální regrese navržená v r. 1979 Clevelandem. V práci [1] je uveden algoritmus této metody i jeho modifikace pro případ, že rozptyl náhodných fluktuací ε_i, $i = 1, \ldots, n$ není konstantní, ale je funkcí času $\sigma^2(t)$.

Tato Clevelandova práce společně s nepublikovanou prací Matyasovszkého [2] byla podnětem k sestavení programu v jazyce Turbo Pascal 5.5, který vypočte robustní vážené lokální odhady neznámé trendové funkce a graficky je znázorní. Tento program umožňuje načíst n pozorování časové řady, zvolit jednu ze tří váhových funkcí:

$G_1(x) = (1 - |x|^3)^3$ pro $|x| < 1$,
$G_2(x) = \frac{15}{16}(1 - 2x^2 + x^4)$ pro $|x| < 1$,
$G_3(x) = \frac{35}{32}(1 - 3x^2 + 3x^4 = x^6)$ pro $|x| < 1$,
$G_1(x) = G_2(x) = G_3(x) = 0$ pro $|x| \geqslant 1$,

zadat počet opakování k rekurentního postupu, vybrat stupeň m aproximujícího polynomu ($m = 0, 1, 2$) a zvolit šířku regrese $b \in (0, 1)$.

3. Popis simulačního experimentu

Vliv parametrů b, k, m a váhové funkce $G_i(x)$, $i = 1, 2, 3$ na odhad neznámého trendu byl posouzen pomocí simulačního experimentu s trendovou funkcí modelu (1) tvaru

$$f(t) = 0,01(t - 1) + \sin 0,04\pi(t - 1) ,$$

kde proměnná t nabývala hodnot $1, 2, \ldots, 100$. Volba trendové funkce byla ovlivněna skutečností, že v teplotní časové řadě lze očekávat nejenom lineární trend (v tomto případě představovaný složkou $0,01(t - 1)$), ale též periodicitu (reprezentovanou složkou $\sin 0,04\pi(t - 1)$). Pomocí této trendové funkce bylo vytvořeno několik simulovaných časových řad. Střední hodnota náhodných fluktuací byla nulová, rozptyl byl buď konstantní nebo funkcí času, rozdělení náhodných fluktuací rovnoměrné, exponenciální nebo normální.

Pro rovnoměrné rozdělení náhodných fluktuací je charakteristické, že hodnoty časové řady se vyskytují uvnitř daného intervalu se stejnou pravděpodobností jako uvnitř jiného intervalu stejné délky. Do časové řady s tímto rozdělením náhodných fluktuací byly záměrně vneseny hrubé chyby v okolí lokálních extrémů trendové funkce, aby bylo možno zkoumat vliv odlehlých pozorování na odhad trendu. Tyto chybné hodnoty jsou na obr. 5 označeny kroužkem.

V případě exponenciálního rozdělení náhodných fluktuací, které je nesymetrické, se objevují značně extrémní hodnoty (viz obr. 1, 3).

Normální (Gaussovo) rozdělení náhodných fluktuací je symetrické, extrémní hodnoty se vyskytují s malou pravděpodobností (viz. obr. 2, 4, 6).

V simulačním experimentu se též sledovalo porušení předpokladu o konstantním rozptylu náhodných fluktuací a jeho vliv na odhad trendu. Rozptyl normálně rozdělených náhodných fluktuací časové řady byl generován pomocí následujících funkcí:

$\sigma^2(t) = \sin^2 0,04\pi(t - 1)$; vykazuje cyklický trend s největšími hodnotami v okolí lokálních extrémů trendové funkce (obr. 6a);

$\sigma^2(t) = \cos^2 0,04\pi(t - 1)$; vykazuje cyklický trend s nejmenšími hodnotami v okolí lokálních extrémů trendové funkce (obr. 6b);

$\sigma^2(t) = 0,1[0,02\pi(t - 51)]^2$; vykazuje parabolický trend s největšími hodnotami v okrajových bodech časové řady (obr. 6c);

$\sigma^2(t) = 1/\{0,01 + [0,04\pi(t - 51)]^2\}$; vykazuje hyperbolický trend s nejmenšími hodnotami v okrajových bodech časové řady (obr. 6d).

4. Závěry simulačního experimentu

S rostoucí šířkou regrese b se odhad trendu vyhlazuje (obr. 1)

Rozdíly mezi aproximujícími polynomy nultého a prvního stupně jsou nevýznamné, více se liší výsledky získané pomocí polynomu druhého stupně (obr. 2).

S rostoucím počtem opakování k rekurentního postupu se odhad trendu vyhlazuje (obr. 3).

Různá volba váhových funkcí neovlivňuje výrazněji odhad trendu, nejmenší součet čtverců odchylek dává váhová funkce $G_3(x)$ (obr. 4).

Odlehlá pozorování ovlivní odhad trendu jen nepatrně (obr. 5). Robustnost této metody zabraňuje zkreslení, která způsobují odlehlá pozorování. Je to výhodné, pokud se jedná o chybná pozorování, avšak v případě, že odlehlá pozorování mají své fyzikální opodstatnění, nemusí jít vždy o vítaný efekt.

Metoda se dobře přizpůsobuje měnícímu se rozptylu (obr. 6). Tato vlastnost se uplatní zvláště při odhadu trendu těch časových řad, u nichž je narušena homogenita v rozptylu.

Vzhledem ke skutečnosti, že robustní vážená lokální regrese byla použita k odhadu trendu časových řad s rovnoměrným, exponenciálním a normálním rozdělením náhodných fluktuací, můžeme zkoumat též vliv různých typů rozdělení náhodných fluktuací na odhad trendu. Simulační experiment prokázal, že metoda není citlivá na typ rozdělení náhodných fluktuací.

5. Příklad použití robustní vážené lokální regrese k analýze řady průměrných ročních teplot Prahy–Klementina 1771-1990

Ke zpracování této teplotní řady byly zvoleny konstantní parametry $k = 2$, $m = 1$, $G_1(x)$ a proměnlivé b (postupně $b = 0,02,\ 0,03,\ 0,05$). Původní hodnoty analyzované řady s odpovídajícími odhadnutými trendovými funkcemi pro různé hodnoty b jsou znázorněny na obr. 7. V souladu se závěry 4. kapitoly se zdá nejpodstatnější šířka regrese, která ovlivňuje dílčí teplotní maxima a minima postižitelná průběhem trendové funkce. Na tomto parametru tak závisí možný stupeň detailizace popisu na základě získaných trendových funkcí. Podrobný rozbor této teplotní řady je uveden v [3].

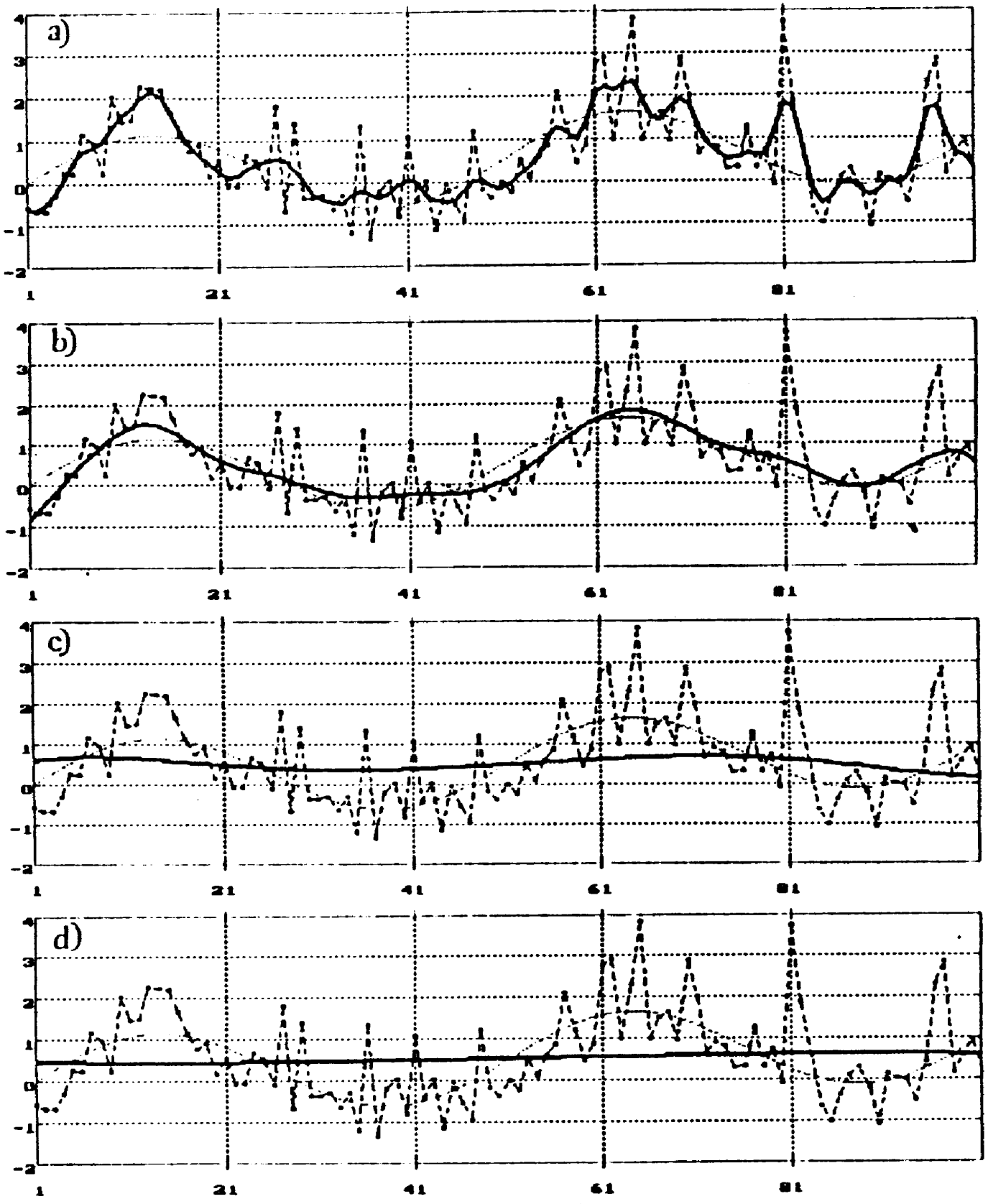
Literatura:

[1] Cleveland, W. S., *Robust locally weighted regression and smoothing scatterplots*, J. Amer. Stat. Assoc., 74, 1979, č. 368, s. 829-836.

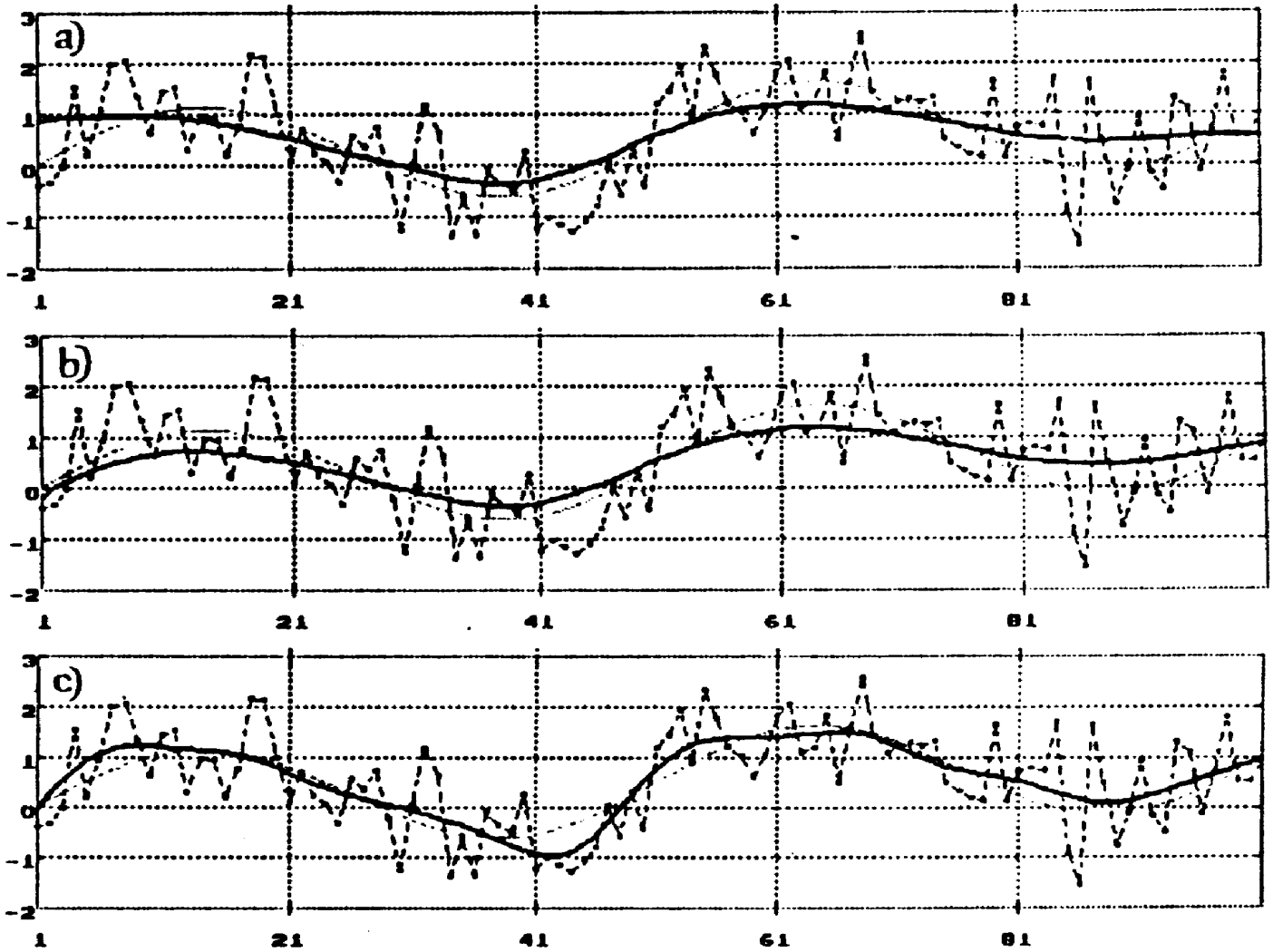
[2] Matyasovszky, J., *Nonparametric methods for trend estimation of climatological time series: a methodological review*, Rukopis (Theor. Appl. Clim.), 1990, nepublikováno.

[3] Michálek, J. - Budíková, M. - Brázdil, R., *Metody odhadu trendu časové řady na příkladu středoevropských teplotních řad*, svazek 9 publikací NKP, Praha 1993, v tisku.

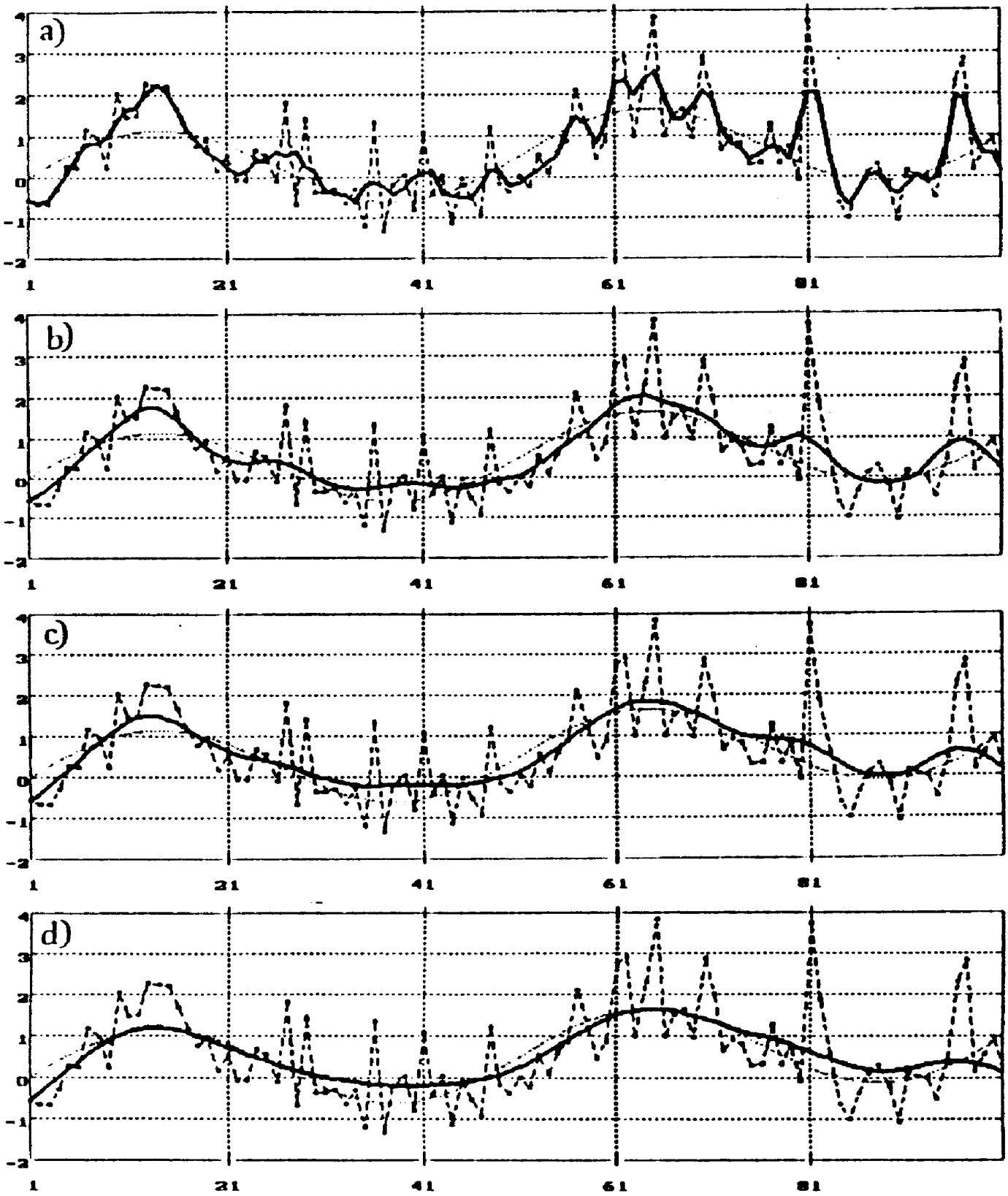
[4] Solow, A. R., *Detecting changes through time in the variance of a long-term hemispheric temperature record: an application of robust locally weighted regression*, J. Climate, 1, 1988, č. 3, s. 290-296.



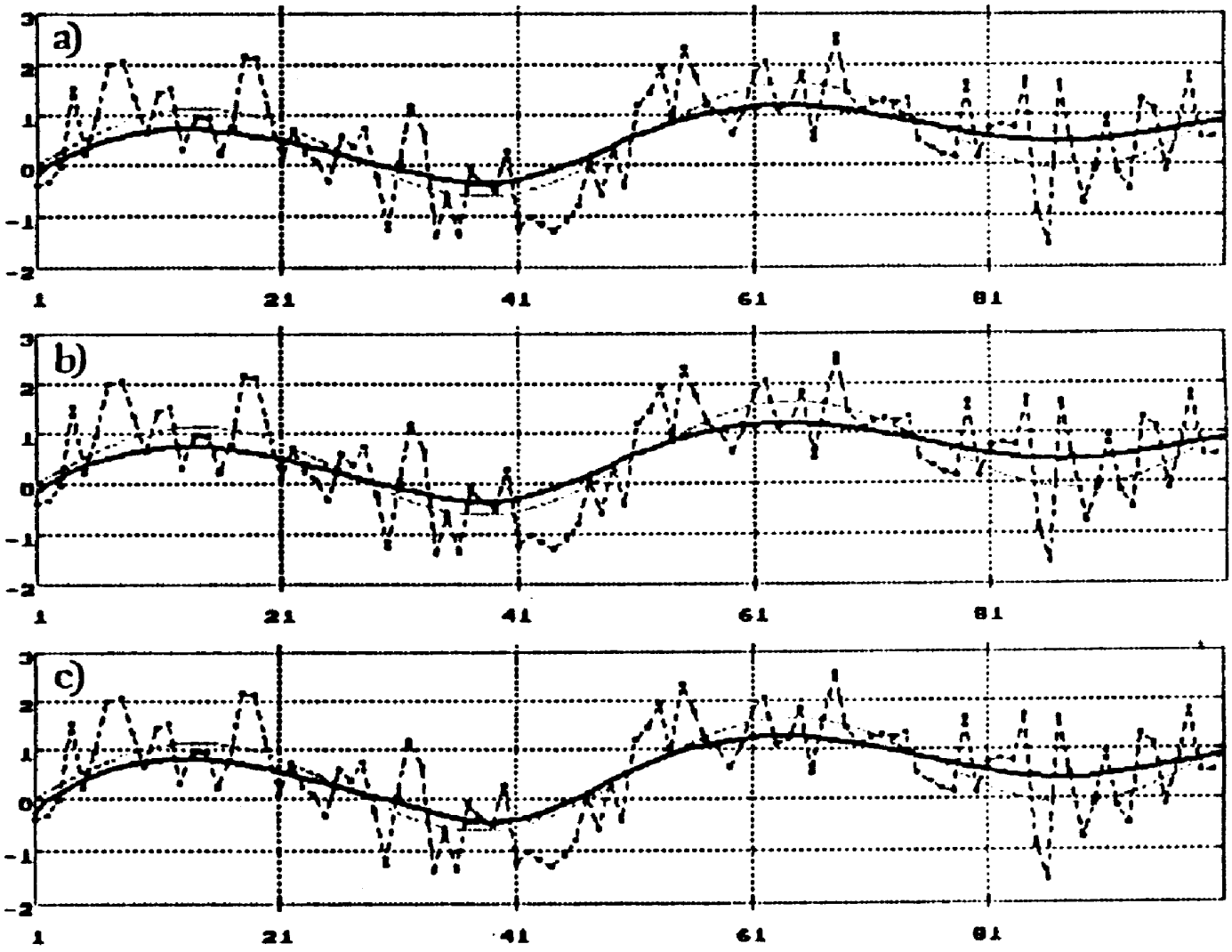
Obr. 1 **Vliv šířky regrese b na odhad trendové funkce simulované řady.** Parametry výpočtu: $k = 1$, $m = 1$, $G_1(x)$



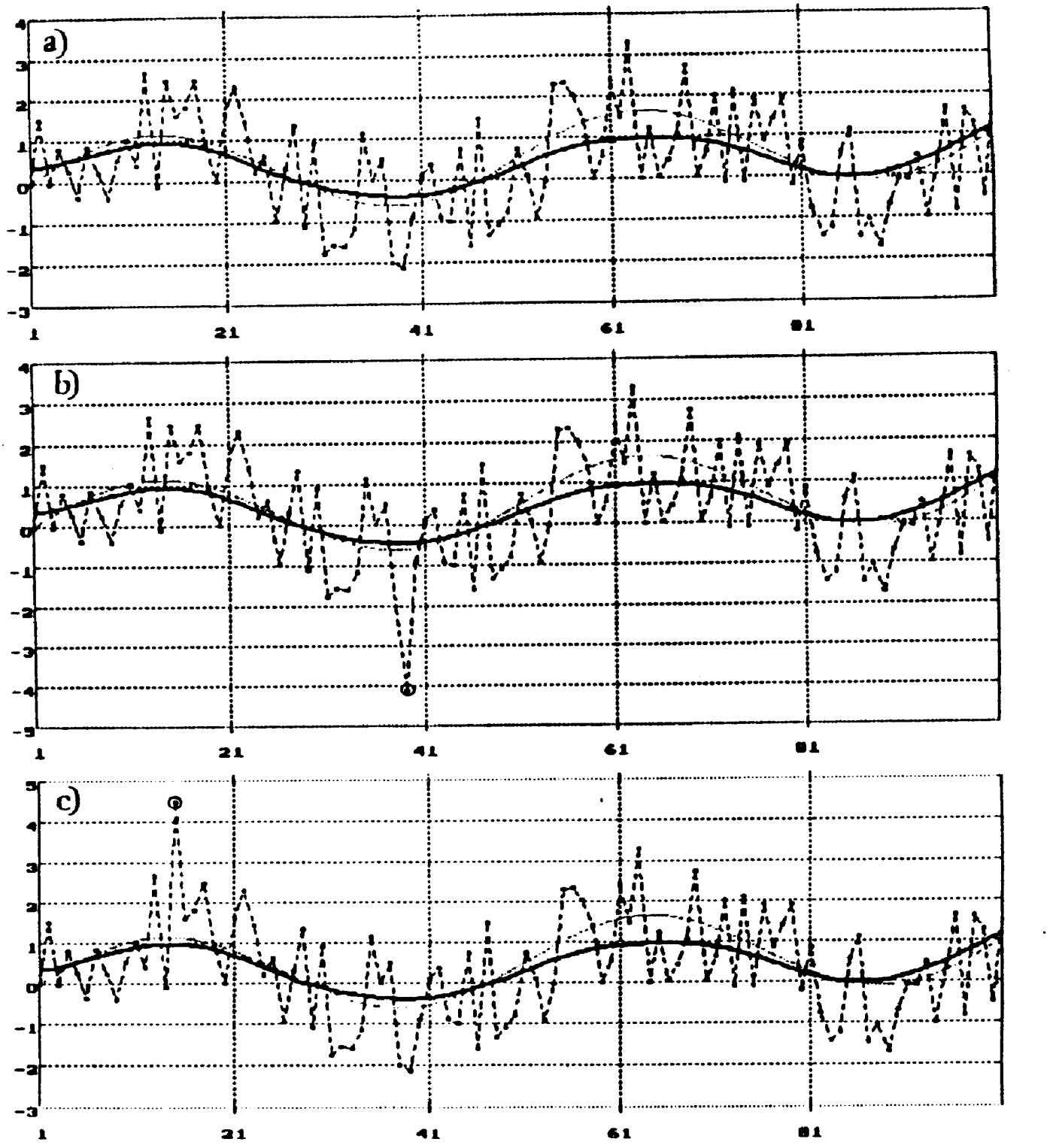
Obr. 2 Vliv stupně aproximujícího polynomu m na odhad trendové funkce simulované řady.
Parametry výpočtu: $k = 2$, $b = 0,1$, $G_1(x)$
a) $m = 0$, b) $m = 1$, c) $m = 2$



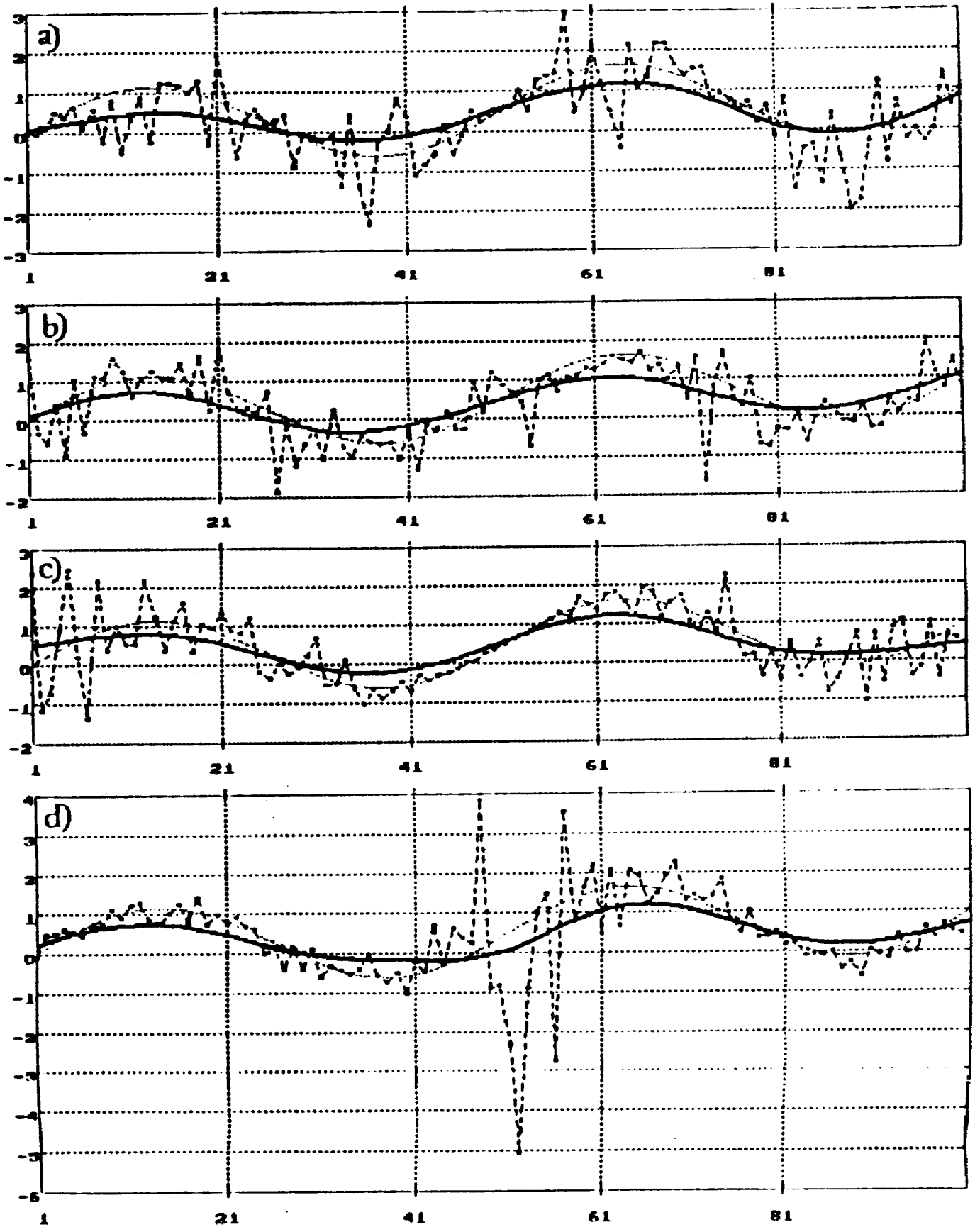
Obr. 3 **Vliv počtu opakování k na odhad trendové funkce simulované řady. Parametry výpočtu: $h = 0,02$, $m = 1$, $G_1(x)$**
$k = 1$, b) $k = 5$, c) $k = 10$, d) $k = 20$



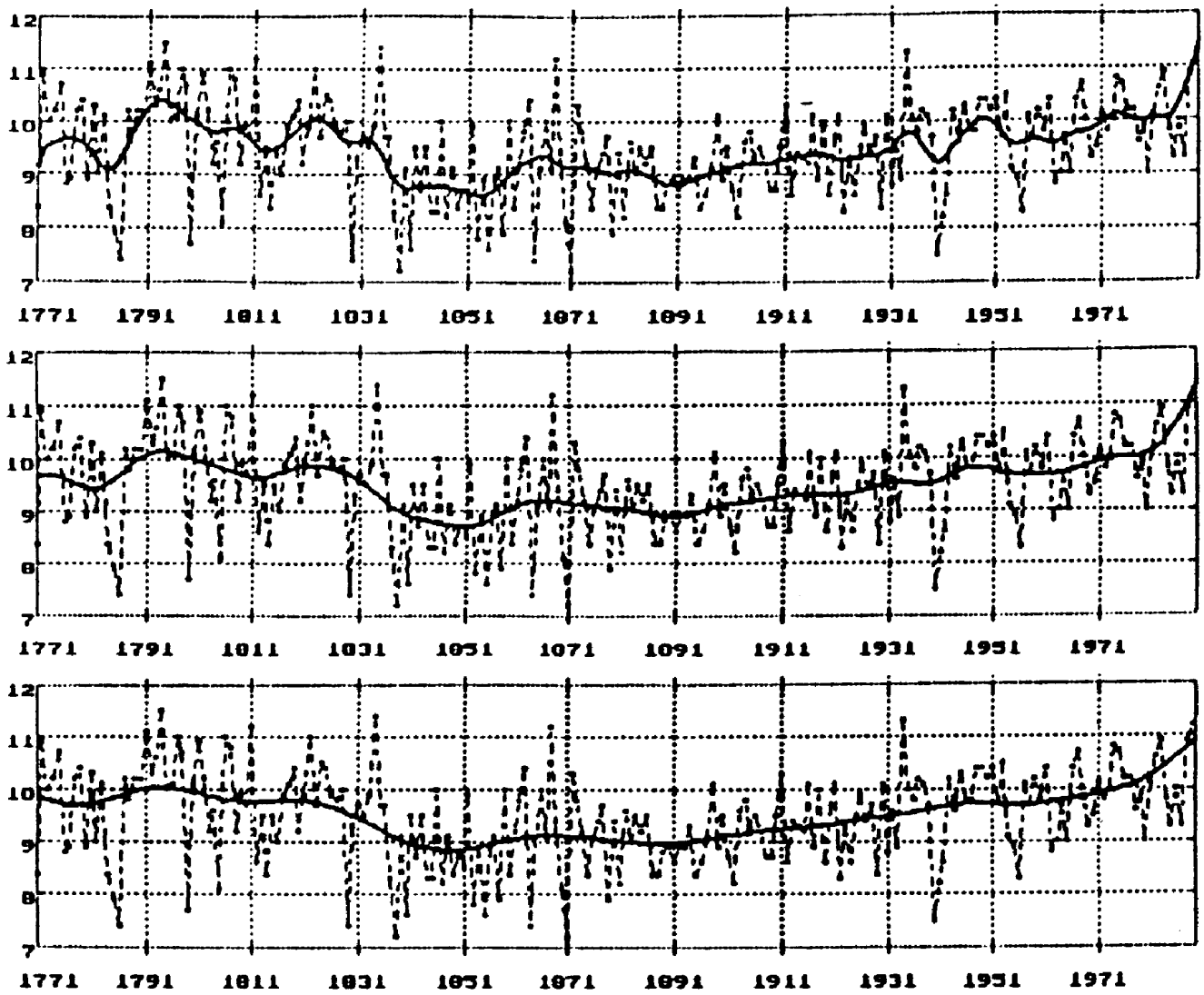
Obr. 4 Vliv váhových funkcí $G_i(x)$, $i = 1,2,3$ na odhad trendové funkce simulované řady.
Parametry výpočtu: $b = 0,1$, $k = 2$, $m = 1$



Obr. 5 **Vliv odlehlých pozorování na odhad trendové funkce simulované řady.** Parametry výpočtu: $b = 0,1$, $k = 1$, $m = 1$, $G_1(x)$
(Chybné hodnoty označeny kroužkem)



Obr. 6 Vliv měnícího se rozptylu normálně rozdělených náhodných fluktuací na odhad trendové funkce simulované řady.
Parametry výpočtu: $h = 0,1$, $k = 2$, $m = 1$, $G_1(x)$



Obr. 7 Chod průměrných ročních teplot vzduchu a trendových funkcí stanovených robustní váženou lokální regresí pro Prahu-Klementinum v období 1771 - 1990.
Parametry výpočtu: $k = 2$, $m = 1$, $G_1(x)$
a) $h = 0.02$, b) $h = 0.03$, c) $h = 0.05$